

KNOWLEDGE AND MIND



NORMAN MALCOLM

Photograph courtesy of Norman Kretzmann

Knowledge
and
Mind
PHILOSOPHICAL ESSAYS

Edited by
Carl Ginet
and
Sydney Shoemaker

New York Oxford
OXFORD UNIVERSITY PRESS
1983

Copyright © 1983 by Oxford University Press, Inc.

Library of Congress Cataloging in Publication Data
Main entry under title:

Y 11 C 35 Knowledge and mind. 1911-1982.

Essays honoring Norman Malcolm.

"Bibliography of Norman Malcolm's writings": p.

Includes index.

1. Philosophy—Addresses, essays, lectures.

2. Malcolm, Norman, 1911- I. Ginet, Carl.

II. Shoemaker, Sydney. III. Malcolm, Norman, 1911-

B29.K59 121 81-22577

ISBN 0-19-503148-2 AACR2

\$29.50

"The Objective Self" copyright © 1980 by Thomas Nagel

Printing (last digit): 9 8 7 6 5 4 3 2 1

Printed in the United States of America

CONTRIBUTORS

G.E.M. ANSCOMBE
Cambridge University

JOHN V. CANFIELD
University of Toronto

JOHN W. COOK
Santa Barbara, California

KEITH S. DONNELLAN
University of California,
Los Angeles

PETER GEACH
Cambridge, England

CARL GINET
Cornell University

BRUCE GOLDBERG
University of Maryland/
Baltimore County

HIDE ISHIGURO
Barnard College

THOMAS NAGEL
New York University

DAVID H. SANFORD
Duke University

SYDNEY SHOEMAKER
Cornell University

GEORG HENRIK VON WRIGHT
Helsinki, Finland

Preface

Over the past four decades Norman Malcolm has been a vigorous and influential figure in Anglo-American philosophy. For those who were his students at Cornell—where he was a member of the Sage School of Philosophy from 1948 until his retirement in 1978—his seriousness, directness, honesty, insistence on clarity, and distrust of the glib and facile, set a valuable and unforgettable example. The many who have learned from him are not limited to those who have been officially his students. His lucid and provocative writings have enlivened and advanced the discussion of central issues in philosophy of mind, theory of knowledge, philosophy of religion, and the philosophies of Descartes and Wittgenstein.

This volume of essays honors Malcolm on his seventy-second birthday (it was to have been his seventieth, but we missed). The contributors are all friends and admirers of Malcolm, and most of them have also been his students or colleagues. The editors are among the fortunate few who have been both students and colleagues. We join the other contributors in expressing our gratitude to Norman Malcolm for what he has taught us, both about particular philosophical topics and about how to do philosophy.

To give the volume unity of content, we have asked the contributors to write on topics in either theory of knowledge or philosophy of mind, these being two of Malcolm's major areas of interest. A bibliography of Malcolm's writings appears at the end of the volume.

Preface

Over the past four decades Norman Malcolm has been a vigorous and influential figure in Anglo-American philosophy. For those who were his students at Cornell—where he was a member of the Sage School of Philosophy from 1948 until his retirement in 1978—his seriousness, directness, honesty, insistence on clarity, and distrust of the glib and facile, set a valuable and unforgettable example. The many who have learned from him are not limited to those who have been officially his students. His lucid and provocative writings have enlivened and advanced the discussion of central issues in philosophy of mind, theory of knowledge, philosophy of religion, and the philosophies of Descartes and Wittgenstein.

This volume of essays honors Malcolm on his seventy-second birthday (it was to have been his seventieth, but we missed). The contributors are all friends and admirers of Malcolm, and most of them have also been his students or colleagues. The editors are among the fortunate few who have been both students and colleagues. We join the other contributors in expressing our gratitude to Norman Malcolm for what he has taught us, both about particular philosophical topics and about how to do philosophy.

To give the volume unity of content, we have asked the contributors to write on topics in either theory of knowledge or philosophy of mind, these being two of Malcolm's major areas of interest. A bibliography of Malcolm's writings appears at the end of the volume.

Ithaca, N.Y.
June 1982

C.G.
S.S.

Contents

Moore and Scepticism	3
<i>John W. Cook</i>	
Justification of Belief: A Primer	26
<i>Carl Ginet</i>	
On Causal Knowledge	50
<i>Georg Henrik von Wright</i>	
Scepticism and Sanity	63
<i>Hide Ishiguro</i>	
Kripke and Putnam on Natural Kind Terms	84
<i>Keith S. Donnellan</i>	
Discovering Essence	105
<i>John V. Canfield</i>	
The Perception of Shape	130
<i>David H. Sanford</i>	
Abstraction Reconsidered	160
<i>Peter Geach</i>	
The Causation of Action	174
<i>G.E.M. Anscombe</i>	
Mechanism and Meaning	191
<i>Bruce Goldberg</i>	

The Objective Self	211
<i>Thomas Nagel</i>	
On an Argument for Dualism	233
<i>Sydney Shoemaker</i>	
Books and Articles by Norman Malcolm	259
Index	265

KNOWLEDGE AND MIND

Moore and Scepticism

JOHN W. COOK

G. E. Moore's attempts to defend *Common Sense* against philosophical scepticism and to give a proof of an external world¹ have left many philosophers puzzled. No philosopher I know of is confident that Moore, in this instance, entirely succeeded in doing what he had set out to do. Yet there are many who believe that, while Moore's way of replying to the sceptic was not entirely satisfactory, there was nevertheless something right—and something important—in his procedure, and accordingly they have attempted in various ways to reconstruct or reformulate that part which they believe to be right. I do not share this view. I believe that Moore was entirely wrong in replying to the sceptic as he did and that no part or aspect of his procedure can be successfully salvaged or defended. In this essay I will set forth some of my reasons for taking this dissenting view.

I

A striking feature of Moore's refutations of scepticism regarding the external world is that he seems not to have thought it essential to begin by examining the arguments which have

1. "A Defence of Common Sense" in *Contemporary British Philosophy, Second Series*, ed. J. H. Muirhead (George Allen and Unwin: London, 1925), and "Proof of an External World," *Proceedings of the British Academy*, 23 (1937). Both papers are reprinted in Moore's *Philosophical Papers* (George Allen and Unwin: London, 1959), and my page references will be to this volume.

traditionally led to such scepticism. Although he did, from time to time, turn his attention to those arguments, his attitude in this matter seems to have remained that which he summed up in the following remarks about the views of sceptical philosophers:

. . . it seems to me a sufficient refutation of such views as these, simply to point to cases in which we do know such things [as the sceptic denies we can know]. This, after all, you know, really is a finger: there is no doubt about it: I know it, and you all know it. And I think we may safely challenge any philosopher to bring forward any argument in favor either of the proposition that we do not know it, or of the proposition that it is not true, which does not at some point, rest upon some premise which is, beyond comparison, less certain than is the proposition which it is designed to attack. The questions whether we do ever know such things as these, and whether there are any material objects, seem to me, therefore, to be questions which there is no need to take seriously: they are questions which it is quite easy to answer, with certainty, in the affirmative.²

Moore's view, then, is that one can be entirely confident in advance of examining any such sceptical argument that it contains some defect.

His confidence on this point was born of the conviction that the sceptic's conclusion is vulnerable to a direct attack, an attack which can be seen to succeed without our giving prior attention to the sceptic's argument. And how does Moore attack the conclusion? In the passage quoted above he says that we can "simply point to cases in which we do know such things"—such things as the sceptic denies we can know. This does not, however, mean that Moore replies to the conclusion that one cannot know whether there is an external world by simply retorting, "On the contrary, I do know there's an external world." Rather, to arrive at "cases" which will serve his purpose, Moore employed a technique which he once called "translating into the concrete,"³ a technique by which he transformed highly general philosophical claims into "concrete" instances or cases.

2. "Some Judgments of Perception," reprinted in *Philosophical Studies* (Humanities Press: New York, 1951), p. 228.

3. "The Conception of Reality," reprinted in *ibid.*, p. 209.

Thus, the sceptic's conclusion that we have no knowledge of an external world is transformed into some such concrete claim as that one cannot know whether this is a finger or that a person cannot know whether he has hands. This, then, enables Moore to make his reply by saying, for example, "But this is a finger, and we all know that it is" or "But I do have two hands, and this is something I know to be so."

This technique of translating into the concrete is one that Moore used in dealing with a variety of philosophical issues, and one of the justifications that he gave for its use was that it served to diminish the plausibility a philosophical claim may have when it is stated in a highly general form. "So long as [the philosophical view in question] is merely presented in vague phrases," said Moore, "it does in fact sound very plausible. But as soon as you realize what it means in particular instances. . . it seems to me to lose all its plausibility."⁴ Now it could hardly be doubted that this is the effect that Moore means to achieve by translating into concrete instances the sceptic's conclusion about our knowledge of the external world. For so long as we are confronted with the question "Is there an external world?" it may seem at least plausible for the sceptic to say, "No one can know," whereas if we put to ourselves the question, "Do I have hands?" one may feel impelled to answer, "Of course I do! How could I ever doubt such a thing?" Thus, whereas it would have seemed merely perverse had Moore flown in the face of the sceptic's argument by retorting, "But there is an external world. I *know* that there is," it seems not at all perverse, but highly impressive, when he replies by insisting that he knows he has hands. He seems to bring us back to reality, like someone who dismisses foolish speculations with the admonition, "Stop talking nonsense and get down to cases!"

Yet, impressive though this may be, I have misgivings about it. I cannot help but feel that Moore is somehow dealing unfairly with the sceptic when he translates into the concrete. I want, therefore, to examine this move carefully, and I will do so in two stages. I will begin by trying to make clear exactly

4. *Some Main Problems of Philosophy* (George Allen and Unwin: London, 1953), p. 135.

what is at issue between Moore and the sceptic, and here my concern will be to show that both Moore and the sceptic understand the issue to be a metaphysical one. I will then go on to consider what is achieved by Moore's translating into the concrete.

II

It is often easy to forget, while reading Moore's essays, that this concern with scepticism was a metaphysical one. It is easy to forget this because, for the most part, Moore keeps our attention riveted upon his own "concrete cases," such as "This is a finger" or "I am dressed and not absolutely naked." But we will have misunderstood Moore's aims if we understand him to have been, at any point, wholly or chiefly concerned with no broader issue than whether, for example, he was dressed or naked. His real interest, of course, was not in any such question at all. His real interest here was that which he described as follows in his 1910-1911 lectures:

. . . the most important and interesting thing which philosophers have tried to do is no less than this; namely: To give a general description of the whole Universe, mentioning all of the most important kinds of things which we *know* to be in it, considering how far it is likely that there are in it important kinds of things which we do not absolutely *know* to be in it, and also considering the most important ways in which these various kinds of things are related to one another.⁵

Moore goes on to say that providing such an account of the Universe is "the first and most important problem of philosophy." Now it could hardly be doubted that it was this problem to which Moore was addressing himself in formulating his defense of Common Sense and in giving his proof of an external world. That is to say, Moore was engaged in good old, unabashed metaphysics. There are numerous indications that this was his concern, and I will now mention a few of these.

First, Moore was concerned to defend Common Sense and to give a proof of an external world because other philosophers

5. *Ibid.*, p. 1.

had advanced arguments, such as those in Descartes's First Meditation, which purport to cast doubt upon the existence of an external world of material objects. Or to put the matter in another way, Moore was concerned to reply to that form of *scepticism which, as he himself explains it, "asserts that we simply do not know at all whether there are any material objects in the Universe at all. It admits that there may be such objects; but it says that none of us knows that there are any."*⁶ In what follows I will occasionally refer to this as "Cartesian scepticism," in order to emphasize that the scepticism in question holds that, for all we know, there may be no external world of material objects. Such scepticism is quite different from that of some other sceptics who, perhaps because they are materialists, propose no doubts about an 'external world' but who hold that no contingent proposition can be known to be true. The latter is a purely epistemological scepticism, whereas the former, which was of great interest to Moore, has plainly metaphysical implications. It tells us that we cannot know what kind of Universe we inhabit.

Second, as is implicitly conceded by Moore in a passage quoted above, he regards it as a genuine question, albeit an easily answered one, whether there are any material objects in the Universe. That he so regarded the question is due, or partly due, I think, to something that Arthur Murphy was pointing out when he remarked that "Moore rejects the sceptical conclusion, but he seems, at least at times, to have retained the assumption from which it was naturally derived."⁷ What Murphy was alluding to is the fact that Moore did not reject the representative theory of perception, with its notions of direct perception and sense impressions (or sense-data). And I take it that Murphy is suggesting, although he does not explicitly say this, that so long as Moore was prepared to accept these features of the sceptic's own view, he was in no position to claim to know such things as the sceptic denies he can know. Moore, of course, was far from ready to agree with this; he thought it possible to reconcile our having knowledge of the external

6. *Ibid.*, p. 19.

7. "Moore's 'Defence of Common Sense,'" in *The Philosophy of G. E. Moore*, ed. Paul Schilpp (Tudor: New York, 1942), p. 316.

world with the representative theory of perception. Nevertheless, he did acknowledge that the analysis which he found most plausible for such judgments of perception as "I see two coins" was an analysis which posed a problem, for given this analysis, "it is difficult," he said, "to answer the questions: How can I ever come to know that these sensibles [sense-data] have a 'source' at all? And how do I know that these 'sources' are circular?"⁸ The view which he eventually came to hold was that our knowledge of the external world is "based on an analogical or inductive argument," so that one crucial point on which he differed from the sceptic, he said, is that "I am inclined to think that what is 'based on' an analogical or inductive argument, in the sense in which my knowledge or belief that this is a pencil is so, may nevertheless be certain knowledge and *not* merely more or less probable belief."⁹

I will not here enter into the question whether Moore could in this way reconcile the representative theory of perception with our having knowledge of the external world. I do not think that he could. But here I want only to point out that Moore, along with the sceptic, understood the question about our knowledge of an external world to be a question about whether we have knowledge of 'transcendental' objects. That is to say, in answering the question "Is there an external world?" he understood himself to be answering some such question as "Is there anything, such as a pair of coins or a pair of hands, on the far side of sense-data?"

My reason for insisting on this is that it seems to be forgotten or neglected in some discussions of Moore. For example, Wittgenstein seems to have been either forgetting or neglecting this aspect of Moore's thinking when, in *On Certainty*, he suggested that Moore's saying that he knew he had hands was "nonsense" because Moore's saying this lacked a suitable context, a context in which there was, in fact, something to be known.¹⁰ Had Wittgenstein been mindful of the 'transcendental' character of Moore's intended knowledge claim, he would not,

8. "The Status of Sense-data," reprinted in *Philosophical Studies*, p. 196.

9. "Four Forms of Scepticism," in *Philosophical Papers*, pp. 225-26.

10. Ludwig Wittgenstein, *On Certainty* (Blackwell: Oxford, 1969), sections 10, 347, 348, 412, 423, 461, and 464.

I believe, have thought that such a criticism was the relevant one to make. For as Moore understood the matter, there *was* something to be known here, namely, whether one's sense-data have a 'source' (or a 'source' of the kind one takes them to have).

Third, an important feature of Cartesian scepticism is that it conjures into existence a body of hitherto unheard of 'propositions'—metaphysical propositions which it attaches to the things we have occasion to say in our ordinary, non-philosophical discourses. It does this by alleging that many of the things we ordinarily say can be false in a way that never occurs to us in the ordinary affairs of life. For example, if a child were to say to his mother, "I've washed my hands," this, according to the sceptic, can be false not only in the way we all recognize (as when the child is telling a falsehood and still has unwashed hands), but also in the following way: if there is no external world of material objects, so that the one who says (or 'says') that he has washed his hands is a disembodied being, then what he has said (or 'said') is false, not because his hands remain unwashed, but because he has no body and so has no hands at all, neither washed nor unwashed hands. Accordingly, the sceptic maintains that when, ordinarily, one says (or 'says'), "I've washed my hands," this implies the further proposition, "At least two human hands exist" or, more generally, "Material objects exist." Similarly, if someone says, "I am taking a bath," this, according to the sceptic, implies the further proposition "I have a body." Now these allegedly implied propositions are not among the things we ordinarily have occasion to say. Indeed, one cannot even begin to understand them apart from an understanding of Cartesian scepticism. They are plainly metaphysical propositions. But it is propositions of this sort, propositions which are said to be implied by what one might say outside of philosophy, whose truth the sceptic is chiefly concerned to deny we can know.

Now Moore was in complete agreement with this view of language. Thus, although he keeps our attention focused, for the most part, on his concrete cases, we are meant to understand these as having metaphysical implications. Moore makes this explicit in his essay, "Certainty," which begins with his

making seven "assertions," such as "I am standing up" and "I have clothes on," all of which, he later tells us, were "propositions which implied the existence of an *external world*—that is to say, of a world *external to my mind*," so that

if I did know any one of them to be true, when I asserted it, the existence of an external world was at that time absolutely certain. If, on the other hand, as some philosophers have maintained, the existence of an external world is never absolutely certain, then it follows that I cannot have known any one of these seven propositions to be true.¹¹

Accordingly, when we find Moore telling us that he knows, for example, that he has hands or has clothes on, we must bear in mind that on his view these knowledge claims have built into them, by propositional implication, the further metaphysical knowledge claim that he knows there are material objects on the far side of sense-data. This is why, on Moore's view, a wartime casualty who wanted to know whether he still had hands would need to do more than merely probe the bandages or peer beneath them—why, that is, he would need to engage in "an analogical or inductive argument."

Fourth, when Moore undertakes to identify what he calls "the Common Sense view of the world," he does so (in part, at least) by means of metaphysical propositions of the kind mentioned above. When he begins his list of "truisms" by saying, "There exists at present a living human body, which is *my body*," Moore surely did not think of this as something a plain man would ever say, but he did undoubtedly think of it as a metaphysical proposition (one which would contradict "Moore is a disembodied being") and as a proposition which would be *implied* if he were to say to his wife, "I'm taking a bath," or "I've washed my hands." And there are other of his truisms which apparently fall into this category, as when he goes on to say that his body "has existed continuously" since it was born and that it has always been "at various distances" from other things "having shape and size in three dimensions."¹² These truisms, too, I take it, are to be thought of, not as things a plain

11. In *Philosophical Papers*, pp. 242-43.

12. "A Defence of Common Sense," p. 33.

man would ever say, but as metaphysical propositions which are implied by things the plain man says.

Thus, what Moore calls "the Common Sense view of the world" is, in part at least, a plainly metaphysical view, and Moore, in defending Common Sense, meant to be defending this metaphysical view.

I have now mentioned four ways in which Moore, despite his focusing our attention on concrete cases or instances, can be seen to be concerned with metaphysics. There may be additional indications of this in his writings, but I believe that these four are sufficient to make the point inescapable.

What I have wanted to emphasize here can be further elaborated by alluding to Russell's attitude toward the change brought about in philosophy by Wittgenstein's later work, a change which Russell plainly did not understand. His attitude is succinctly expressed in the following passage:

. . . the new philosophy seems to me to have abandoned, without necessity, that grave and important task which philosophy throughout the ages has hitherto pursued. Philosophers from Thales onwards have tried to understand the world. . . . I cannot feel that the new philosophy is carrying on this tradition. It seems to concern itself, not with the world and our relation to it, but only with the different ways in which silly people can say silly things. If this is all that philosophy has to offer, I cannot think that it is a worthy subject of study.¹³

Now while Russell was certainly not in sympathy with Moore's defense of Common Sense or his proof of an external world, he would not have thought that in these endeavors Moore had abandoned the metaphysical aims of traditional philosophy. And in this I would agree with Russell, for I want to insist that Moore was engaged in a metaphysical enterprise to which the thrust of Wittgenstein's later work is fundamentally opposed. Granted that Moore, like Wittgenstein, would have rejected Russell's characterization of what people say outside of philosophy as being "silly," nevertheless Moore was in agreement with Russell (and in disagreement with Wittgenstein) on a most

13. Bertrand Russell, *My Philosophical Development* (George Allen and Unwin: London, 1959), p. 250.

making seven "assertions," such as "I am standing up" and "I have clothes on," all of which, he later tells us, were "propositions which implied the existence of an *external world*—that is to say, of a world *external to my mind*," so that

if I did know any one of them to be true, when I asserted it, the existence of an external world was at that time absolutely certain. If, on the other hand, as some philosophers have maintained, the existence of an external world is never absolutely certain, then it follows that I cannot have known any one of these seven propositions to be true.¹¹

Accordingly, when we find Moore telling us that he knows, for example, that he has hands or has clothes on, we must bear in mind that on his view these knowledge claims have built into them, by propositional implication, the further metaphysical knowledge claim that he knows there are material objects on the far side of sense-data. This is why, on Moore's view, a wartime casualty who wanted to know whether he still had hands would need to do more than merely probe the bandages or peer beneath them—why, that is, he would need to engage in "an analogical or inductive argument."

Fourth, when Moore undertakes to identify what he calls "the Common Sense view of the world," he does so (in part, at least) by means of metaphysical propositions of the kind mentioned above. When he begins his list of "truisms" by saying, "There exists at present a living human body, which is *my* body," Moore surely did not think of this as something a plain man would ever say, but he did undoubtedly think of it as a metaphysical proposition (one which would contradict "Moore is a disembodied being") and as a proposition which would be implied if he were to say to his wife, "I'm taking a bath," or "I've washed my hands." And there are other of his truisms which apparently fall into this category, as when he goes on to say that his body "has existed continuously" since it was born and that it has always been "at various distances" from other things "having shape and size in three dimensions."¹² These truisms, too, I take it, are to be thought of, not as things a plain

11. In *Philosophical Papers*, pp. 242-43.

12. "A Defence of Common Sense," p. 33.

man would ever say, but as metaphysical propositions which are implied by things the plain man says.

Thus, what Moore calls "the Common Sense view of the world" is, in part at least, a plainly metaphysical view, and Moore, in defending Common Sense, meant to be defending this metaphysical view.

I have now mentioned four ways in which Moore, despite his focusing our attention on concrete cases or instances, can be seen to be concerned with metaphysics. There may be additional indications of this in his writings, but I believe that these four are sufficient to make the point inescapable.

What I have wanted to emphasize here can be further elaborated by alluding to Russell's attitude toward the change brought about in philosophy by Wittgenstein's later work, a change which Russell plainly did not understand. His attitude is succinctly expressed in the following passage:

. . . the new philosophy seems to me to have abandoned, without necessity, that grave and important task which philosophy throughout the ages has hitherto pursued. Philosophers from Thales onwards have tried to understand the world. . . . I cannot feel that the new philosophy is carrying on this tradition. It seems to concern itself, not with the world and our relation to it, but only with the different ways in which silly people can say silly things. If this is all that philosophy has to offer, I cannot think that it is a worthy subject of study.¹³

Now while Russell was certainly not in sympathy with Moore's defense of Common Sense or his proof of an external world, he would not have thought that in these endeavors Moore had abandoned the metaphysical aims of traditional philosophy. And in this I would agree with Russell, for I want to insist that Moore was engaged in a metaphysical enterprise to which the thrust of Wittgenstein's later work is fundamentally opposed. Granted that Moore, like Wittgenstein, would have rejected Russell's characterization of what people say outside of philosophy as being "silly," nevertheless Moore was in agreement with Russell (and in disagreement with Wittgenstein) on a most

13. Bertrand Russell, *My Philosophical Development* (George Allen and Unwin: London, 1959), p. 230.

fundamental point, namely, that which Russell once stated as follows: "It is undeniable that our every-day interpretations of perceptive experience, and even all our every-day words, embody theories."¹⁴ By "theories" here Russell meant metaphysical theories, including the 'theory' that there exists an external world of material objects. And Moore, as we have already observed, is in agreement with Russell on this point. Thus, I think it is beyond dispute that Moore, like Russell, thought of "ordinary language" in such a way as to include in it, not only the things plain men actually say, but also the various metaphysical propositions or theories or views of the world that are allegedly implied by what plain men actually say.

It was, of course, this conception of "ordinary language" that led Russell to think that philosophy had taken a disastrous turn when other philosophers came along proposing to settle philosophical questions by appealing to "ordinary language." Given Russell's view of what "ordinary language" includes, namely, questionable metaphysical theories, not only about an external world, but also about personal identity, free will, causation, and so on, he was right to think that "ordinary language" cannot be our touchstone of truth in philosophy—and right, also, to think that when other philosophers appeal to "ordinary language" as though it were such a touchstone, they beg all the important questions. And very likely it was for this same reason (in part, anyway) that Moore so vigorously rejected the suggestion that he, in defending Common Sense, was really, in an obscure way, doing nothing more than recommending that we all speak with the vulgar.¹⁵ Moore very likely recognized that if he were merely recommending that we employ "ordinary language," he would have been begging the question, namely, the question whether the Common Sense theories that are built into "ordinary language" are true. He would no doubt have allowed that in defending Common Sense he was also, by implication, defending ordinary language, but he would have

14. Bertrand Russell, *An Inquiry into Meaning and Truth* (George Allen and Unwin: London, 1951), p. 124.

15. Moore's rejection of this suggestion comes out in his discussion of the attempted reconstructions of his method by Alice Ambrose and Morris Lazero-witz, "A Reply to My Critics" in *The Philosophy of G. E. Moore*, pp. 674-75.

insisted that his way of defending both was by insisting that material objects do exist and that, contrary to the sceptic's view, this is something he knows to be true.

Thus, Moore was in essential agreement with Russell in these matters and differed from him only in the comparatively unimportant respect that whereas Russell remained a sceptic regarding the truth of the Common Sense metaphysics, Moore thought that such scepticism was untenable.

III

The question to which we must eventually address ourselves is how Moore could have thought that he knew something which the sceptic denies one can know. Before turning to this question, however, it may be of some use to make explicit several features of Cartesian scepticism which are not always borne in mind by opponents of such scepticism.

1. The sceptic's conclusion, as Moore expresses it in a passage quoted earlier, is that "we simply do not know at all whether there are any material objects in the Universe at all." And because this is the sceptic's conclusion, anyone who would undertake directly to challenge this conclusion would be begging the question if, in making his challenge, he took it for granted, whether explicitly or implicitly, that he has a body and a normal complement of bodily senses: eyes, ears, etc.

An essential feature of Cartesian scepticism is the sweeping nature of the grounds it adduces for philosophical doubt, namely, its claim that one can never recognize whether one is actually seeing such things as rocks and trees or (as in a dream) only seeming to see such things. This being the strategy of the sceptic's argument, it precludes our replying to the sceptic in anything like the way in which one might answer the sort of scepticism one occasionally encounters in daily life. If, in a case of the latter sort, someone scoffingly said to me, "You didn't know that that gun was loaded," I could sometimes rightly reply, "But I did know it was loaded. I was there and *saw* it being loaded." If, on the other hand, the sceptic, elaborating his position, were to say that one can never know such a thing as that a gun is loaded, one could not rebuff his scepticism in the

aforementioned way, for he would, quite rightly, protest that such an answer begs the whole question by merely assuming, not only that one was on the scene with a full complement of bodily senses, but also that one can successfully distinguish seeing a bullet from seeming to see a bullet. Accordingly, if one would convincingly dispute the sceptic's conclusion without first disputing his premises, it behooves one to avoid depicting that conclusion as though it were vulnerable to anything like an ordinary rebuff.

2. In trying to understand the Cartesian's scepticism about the external world, it would be a mistake to think of it as though it could be the sum of a vast number of 'isolated' doubts about the existence of such particular things as this cottage in which I am now sitting or the desk at which I am now writing. And yet it is easy to make such a mistake. Indeed, such an interpretation might be suggested by Descartes's own words. For as part of his dream argument he says, "Well, suppose I am dreaming, and these particulars, that I open my eyes, shake my head, put out my hand, are incorrect; suppose even that I have no such hand, no such body," and further on he writes: "I will consider myself as having no hands, no eyes, no flesh, no blood, no senses, but just having the false belief that I have all these things."¹⁶ Now one might think that one could extrapolate from these remarks and say that one of the things which Descartes's argument has brought into question is whether Descartes has any hands. But is this true? I want to suggest that it is not.

First of all, Descartes raises no doubt especially about his hands. He did not, for example, harbor a suspicion that there was a ghoulish surgeon on the loose who went about in the night severing people's hands and replacing them with scarcely detectable imitations. Moreover, if ever one were to entertain such a doubt, a doubt about whether he or someone else has lost his hands through accident or amputation, then there are certain other things which could not be in doubt—for example, that the person in question has arms or at least a head and torso. Consider what it would be like, after all, to have such a

16. Descartes, *Philosophical Writings*, eds. E. Anscombe and P. Geach (Nelson: Edinburgh, 1954), pp. 62 and 65.

doubt. Apparently it is, or was, not uncommon for the beggars of India to kidnap children, cut off their hands, and put them back on the street with bandaged stumps, the idea being that a child without hands would make a pitiful spectacle and thus be a more effective beggar. Now if I happen to have heard of this dreadful practice, and a friend of mine who is with me on the streets of Calcutta remarks, pointing to a bandaged child, "Oh, that poor child! What can be wrong with her hands?" I might reply, "I wonder if she even has any hands." And I would then go on to tell about the kidnapping and mutilation of children. Here, then, is a case in which I could be said to be in doubt about whether someone has any hands, but obviously my doubt could not be expanded to include whether the child has arms, a torso, a head. After all, it was only because I saw the child's bandaged arms that I had a doubt in the first place. There might, of course, be a different sort of case, one in which I hear conflicting stories about some accident victim, one story being that he lost his hands in the accident, the other being that he was entirely incinerated in the accident. In that case, I may have various doubts about what happened to the accident victim, but none of these could be expressed by my saying, "He may have been incinerated, so that there are no remains, and also have no hands." This would make no sense unless what I had meant to say was that he first lost his hands in the accident and then was incinerated in the fiery aftermath.

From this we can see that if we were to say that the Cartesian sceptic doubts whether he (or whether Moore) has any hands, we would be implying that the sceptic is willing to concede that he (or that Moore) has arms or at least a head and torso. But the Cartesian sceptic is not, of course, willing to concede such a thing. It would therefore be extremely misleading to say that Descartes doubted whether he had hands. His scepticism regarding the external world will be properly understood only if we understand that it cannot be divided up into subsidiary isolated doubts which themselves suggest that the sceptic concedes the existence of most of our familiar world. While it is true that the sceptic may say, if pressed, that one cannot know whether there is an elm tree growing in front of his house, it would be a misrepresentation to infer from this that his scepti-

cism is comprised, in part, of a doubt about the vegetation in his own front yard, as though he did not gainsay the real estate but only claimed ignorance of what was growing there.

If, then, one is properly to reply to such scepticism, one must avoid representing the sceptic as though he harbored such isolated doubts as those mentioned above.

3. As the previous point has already suggested, there is a problem about how the sceptic's conclusion is to be thought of as impinging on what we say and think outside of philosophy. Consider now a further aspect of this problem.

The sceptic, like everyone else, regularly engages in ordinary discourse and is occasionally faced with answering such questions as "Did you know the gun was loaded?" and "Did you know you left that door open?"—questions which we so understand that a person will, in some circumstances, be lying if he says, "No, I didn't know." Now how is the sceptic to think of his scepticism in relation to such questions as they arise in ordinary discourse? Is he going to endorse a negative answer to such questions, regardless of the circumstances? Will he, in other words, maintain that one should invariably reply in the negative to such a question as "Did you know the gun was loaded?" on the grounds that one can never know whether there is an external world?

Sceptics have not, I think, always made clear their solution to this problem. But one solution (and a solution that some of them have hinted at) would seem to be the following:

It would be a mistake to interject one's philosophical scepticism into the discourse of the plain man (assuming that there are plain men). And the reason it would be a mistake is this. If the one who asks, "Did you know the gun was loaded?" does not mean to be asking a philosophical question, a question which probes philosophical issues, but means only to be asking a question within the framework of the Common Sense view of the world, it would create a highly misleading answer in some cases if one were to say, "No, I didn't know that," while thinking to oneself that of course one *couldn't* know such a thing because the existence of an external world is un-

knowable. So if one is going to answer the question at all, one must answer it in the spirit in which it is asked. And the same goes for such other questions, asked by plain men, as "Could you be mistaken?" and "Do you know where he is living now?" If such a question is put to one without philosophic intent, then the sceptic, like anyone else, must answer it, if he answers at all, with a view to what the inquirer meant to ask. And in general one must not interject one's philosophical scepticism into the midst of ordinary discourse, for to do so would not only create misunderstanding of an ordinary sort but would misrepresent the nature of philosophical scepticism. Cartesian scepticism calls in question the assumptions underlying ordinary discourse and therefore attacks it, not in a piecemeal fashion, but as a whole. The decision facing the sceptic is whether to engage in ordinary discourse at all or whether to give it up altogether. And if he chooses the former alternative, then he must talk like everyone else. As Russell once remarked,¹⁷ he would, if he were "a true philosopher," speak only about sense-data, but "life is too short," he said, and so he speaks as a non-philosopher would.

Some such account, it seems to me, is the right one to give of how Cartesian scepticism impinges on what we say outside philosophy. It finds fault with "ordinary language" as a whole and is not to be thought of as finding philosophical faults within our ordinary discourses.

This point can be put differently as follows. The Cartesian sceptic can readily allow that within ordinary discourse there is a distinction to be observed between circumstances in which one could sensibly raise a question or express a doubt about or plead ignorance of such a thing as whether one has hands and circumstances in which one could not sensibly do so. For example, a wounded soldier who feared that his hands might have to be amputated, and who, upon reviving from anesthesia in the field hospital, had not yet probed the bandages, might ask the nurse, "Do I still have my hands?" or might groggily reply,

17. Bertrand Russell, *Philosophy* (Norton: New York, 1927), pp. 243-44.

cism is comprised, in part, of a doubt about the vegetation in his own front yard, as though he did not gainsay the real estate but only claimed ignorance of what was growing there.

If, then, one is properly to reply to such scepticism, one must avoid representing the sceptic as though he harbored such isolated doubts as those mentioned above.

3. As the previous point has already suggested, there is a problem about how the sceptic's conclusion is to be thought of as impinging on what we say and think outside of philosophy. Consider now a further aspect of this problem.

The sceptic, like everyone else, regularly engages in ordinary discourse and is occasionally faced with answering such questions as "Did you know the gun was loaded?" and "Did you know you left that door open?"—questions which we so understand that a person will, in some circumstances, be lying if he says, "No, I didn't know." Now how is the sceptic to think of his scepticism in relation to such questions as they arise in ordinary discourse? Is he going to endorse a negative answer to such questions, regardless of the circumstances? Will he, in other words, maintain that one should invariably reply in the negative to such a question as "Did you know the gun was loaded?" on the grounds that one can never know whether there is an external world?

Sceptics have not, I think, always made clear their solution to this problem. But one solution (and a solution that some of them have hinted at) would seem to be the following:

It would be a mistake to interject one's philosophical scepticism into the discourse of the plain man (assuming that there are plain men). And the reason it would be a mistake is this. If the one who asks, "Did you know the gun was loaded?" does not mean to be asking a philosophical question, a question which probes philosophical issues, but means only to be asking a question within the framework of the Common Sense view of the world, it would create a highly misleading answer in some cases if one were to say, "No, I didn't know that," while thinking to oneself that of course one *couldn't* know such a thing because the existence of an external world is un-

knowable. So if one is going to answer the question at all, one must answer it in the spirit in which it is asked. And the same goes for such other questions, asked by plain men, as "Could you be mistaken?" and "Do you know where he is living now?" If such a question is put to one without philosophic intent, then the sceptic, like anyone else, must answer it, if he answers at all, with a view to what the inquirer meant to ask. And in general one must not interject one's philosophical scepticism into the midst of ordinary discourse, for to do so would not only create misunderstanding of an ordinary sort but would misrepresent the nature of philosophical scepticism. Cartesian scepticism calls in question the assumptions underlying ordinary discourse and therefore attacks it, not in a piecemeal fashion, but as a whole. The decision facing the sceptic is whether to engage in ordinary discourse at all or whether to give it up altogether. And if he chooses the former alternative, then he must talk like everyone else. As Russell once remarked,¹⁷ he would, if he were "a true philosopher," speak only about sense-data, but "life is too short," he said, and so he speaks as a non-philosopher would.

Some such account, it seems to me, is the right one to give of how Cartesian scepticism impinges on what we say outside philosophy. It finds fault with "ordinary language" as a whole and is not to be thought of as finding philosophical faults within our ordinary discourses.

This point can be put differently as follows. The Cartesian sceptic can readily allow that within ordinary discourse there is a *distinction* to be observed between circumstances in which one could sensibly raise a question or express a doubt about or plead ignorance of such a thing as whether one has hands and circumstances in which one could not sensibly do so. For example, a wounded soldier who feared that his hands might have to be amputated, and who, upon reviving from anaesthesia in the field hospital, had not yet probed the bandages, might ask the nurse, "Do I still have my hands?" or might groggily reply,

17. Bertrand Russell, *Philosophy* (Norton: New York, 1927), pp. 243-44.

if asked, that he didn't know (or wasn't sure) whether the doctors had saved his hands. But some days later, when his bandages have been removed and he is washing his hands, he could not sensibly, or rationally, say such a thing. That there is such a distinction to be recognized within ordinary discourse is a point the sceptic can readily acknowledge. He can allow, that is, that the plain man, except in highly unusual circumstances, would have to be insane to raise questions or have doubts about whether he has hands.

This, I take it, is something like the point that Descartes was making when, prior to introducing his dream argument, he observed:

But although the senses may sometimes deceive us about some minute or remote objects, yet there are many other facts as to which doubt is plainly impossible, although these are gathered from the same source [the senses]: e.g. that I am here sitting by the fire, wearing a winter cloak, holding this paper in my hands, and so on. Again, these hands, and my whole body—how can their existence be denied? Unless indeed I likened myself to some lunatics, whose brains are so upset by persistent melancholy vapours that they firmly assert . . . that they are clad in purple when really they are naked; or that they have a head of pottery, or are pumpkins, or are made of glass; but then they are madmen, and I should appear no less mad if I took them as a precedent for my own case.¹⁸

Having given himself this caution, Descartes goes on immediately to formulate his sceptical argument, and this fact should suggest to us that there must be some relevant difference between the conclusion of that argument and the ravings of a madman, despite any seeming similarities. And one difference, surely, is that the sceptic, while treating as dubious (what he takes to be) the metaphysical assumptions of "ordinary language," is not thereby bound to conduct his non-philosophical conversations differently from any other rational man. It therefore behooves one who would persuade us to reject the sceptic's conclusion to avoid, in his persuasions, any suggestion that the sceptic's conclusion is (or resembles) a form of lunacy.

18. Descartes, *Philosophical Writings*, p. 62.

Now with these various points in mind, let us turn to the question of how Moore could have thought that he knew something which the sceptic denies one can know.

IV

As I remarked earlier, it would have seemed merely perverse had Moore, in attacking the sceptic's conclusion, flatly replied, "But there *are* material objects on the far side of sense-data. I *know* there are!" Yet, although he is presumably aiming to make this same point when he translates into the concrete, it does not seem as though he is being merely perverse when he says, for example, "I know this is a finger" or "I know I have hands." And not only did it not seem perverse of Moore to say this, it seemed, to him at least, as though this were the *right* thing to say. Why?

To find the answer to this question, we must, I think, bear in mind that Moore understood the matter at hand to be this: either I do know that there are material objects on the far side of sense-data or (as the sceptic claims) I do not know this. For Moore the truth lay in one or the other of these alternatives. Now when one begins from an assumption of this sort, there are two quite different ways in which one may attempt to establish in which of the alternatives the truth lies. One way is to argue *for* one of the alternatives; the other way is to argue *against* (or to otherwise discredit) one of the alternatives. Perhaps Descartes could be thought of as proceeding in the former way when, in the Sixth Meditation, he argues that God is no deceiver and so the world is as we judge it to be when we make reasonable judgments. Was this also Moore's way of proceeding? It would seem not, for he certainly makes no attempt to prove that he knows, for example, that this is a finger or that he has hands. It would therefore seem reasonable to suppose that he has adopted the second of the aforementioned methods, that of disproving or discrediting one of the two alternatives, namely, the sceptic's conclusion. At any rate, it is useful to think of Moore in this way. For when he says, as he does in a passage quoted earlier, "This, after all, you know, really is a finger: there is no doubt about it: I know it, and you all know it," he tells us nothing that would establish or convince us that he does

know it, and yet he is, in a subtle way, aiming to discredit the other alternative, namely, the sceptic's claim that he does not know it. He has done this, of course, by translating into the concrete. As was pointed out earlier, Moore thought of this technique as robbing his opponents' claims of their plausibility.

This aspect of Moore's method becomes obvious in his "Proof of an External World," where, having said that he had two hands and that this was something he knew to be so, he went on to say:

How absurd it would be to suggest that I did not know it, but only believed it, and that perhaps it was not the case. You might as well suggest that I do not know that I am standing up and talking—that perhaps after all I'm not, and that it's not quite certain that I am!¹⁹

Even more revealing of this aspect of Moore's method is a passage in his essay "Certainty." He begins the essay by making, as he says, "seven assertions," one of which is, "I have clothes on, and am not absolutely naked." He then says:

And I do not think that I can be justly accused of dogmatism or over-confidence for having asserted these things positively in the way that I did. In the case of some kinds of assertions, and under some circumstances, a man can justly be accused of dogmatism for asserting something positively. But in the case of assertions such as I made, made under the circumstances under which I made them, the charge would be absurd. On the contrary, I should have been guilty of absurdity if, under the circumstances, I had not spoken positively about these things, if I spoke of them at all. Suppose that now, . . . instead of saying "I have got some clothes on," I were to say "I think I've got some clothes on, but it's just possible that I haven't." Would it not sound rather ridiculous for me now, under these circumstances, to say "I *think* I've got some clothes on" or even to say "I not only think I have, I know that it is very likely indeed that I have, but I can't be quite sure"? For some persons, under some circumstances, it might not be at all absurd to express themselves thus doubtfully. . . . But for me, now, in full possession of my senses, it would be quite ridiculous to express myself in this way, because the circumstances are such

as to make it quite obvious that I don't merely think that I have [clothes on], but know that I have.²⁰

Here it is entirely clear that Moore's method is to secure our agreement that he does know, not by proving that he knows, but by depicting as ridiculous any suggestion to the contrary. The effect of translating into the concrete, then, is to make the sceptic's position appear ridiculous or even irrational.²¹ But is this fair to the sceptic?

Let us ask: Is the sceptic really prepared to say (or committed to saying) those things which Moore brands as absurd or ridiculous? Is he, for instance, prepared to say, "Moore, you don't know that you have hands," or committed to saying, when situated as Moore was, "It's possible that I have no clothes on"? The question is ambiguous, and ambiguous in a way that mirrors an ambiguity in Moore's own method. For we need to distinguish the question: "Is the sceptic prepared to say (or committed to saying) such things in the course of philosophical discussion?" and the question: "Is the sceptic prepared to say (or committed to saying) such things in the course of ordinary, non-philosophical discourse?" The answer to the first of these questions is plainly "Yes," while the answer to the second, as we observed in the previous section, is "No," for Cartesian scepticism is to be thought of as attacking "ordinary language" as a whole and not in a piecemeal fashion. It would therefore be misleading to say, without qualification, that the sceptic is prepared to say (or committed to saying) such things.

Now it would also be misleading, and for much the same reason, to claim, without qualification, that it would be absurd (ridiculous, irrational) to say, while having no ordinary grounds for doubt, those things which Moore is concerned with. For again we need to distinguish between: (1) "It would be absurd (ridiculous, irrational) for someone having no ordinary grounds for doubt to say such things in ordinary, non-philosophical discourse," and (2) "It would be absurd (ridiculous, irrational) for a Cartesian sceptic, who had no ordinary grounds for doubt,

20. Pp. 227-28.

21. In another essay Moore says: "I do not think it is *rational* to be as certain of [the premises of a sceptic's argument] as of the proposition that I do know that this is a pencil." "Four Forms of Scepticism," p. 226.

to say such things in the course of expounding his philosophical scepticism." If we take Moore to be saying (1), then everyone, including the sceptic, can entirely agree with him, for if someone, not meaning to philosophize, were to say such things in circumstances in which there are no grounds for a rational man to have a doubt, we would take the person to be a lunatic²² or to be saying something ridiculous to get a laugh. On the other hand, if Moore is saying the second of those things, i.e., that it would be absurd for the sceptic to say, in the course of philosophizing, that he (or Moore) doesn't know he has hands or has clothes on, then what Moore is saying is surely false. Sceptics have been saying such things for centuries, and although others of us may take issue with their reasoning, no one who was following their train of reasoning has thought that these philosophers, when drawing their sceptical conclusion, were (or sounded like) lunatics. And the reason for this is that their sceptical conclusion, understood in its philosophical context, does not sound like a doubt for which the sceptic ought to have, yet lacks, ordinary grounds for doubt. The fact that Moore heard a ring of absurdity in the sceptic's conclusion is due to the fact that, quite unwittingly, he has, by translating into the concrete, misrepresented the sceptic, has made it appear that the sceptic ought to have, but lacks, ordinary grounds for doubt.

This misrepresentation comes about in ways that are not easily recognized or described. I believe, however, that we can identify how it comes about if we focus on the following aspect of the passages quoted above. In the passage from "Proof of an External World" Moore tells us that it would be absurd "to suggest" that he didn't know he had hands, but he says nothing about who is to be thought of as making this suggestion or why he makes it. Similarly, in the passage from "Certainty" Moore asks us to "suppose" that he himself had spoken doubtfully about whether he had clothes on, but he fails to tell us what we are to suppose his reason to have been for speaking thus doubtfully. In both passages, then, he invites us to consider someone's saying something, something which he brands as "absurd" or

22. There is, it seems, such a form of lunacy. See "Doubting Mania" in *Dictionary of Philosophy and Psychology*, ed. James Baldwin (Macmillan: New York, 1925), pp. 296-97.

"ridiculous," but in neither passage does he fill in any details as to why the person whom we are to consider saying these things says them. Or rather he does not do so explicitly. Nevertheless, there is something about both passages which, in a covert way, suggests these further details. For the fact that Moore is taking issue with scepticism leads us to think of him as presenting here a sceptic and what sceptics say. Yet Moore's translating into the concrete somehow creates (I don't say *deliberately*) the opposite impression, the impression that our imagined person is not a philosopher engaged in philosophical discussion but rather someone speaking without philosophical intent, someone engaged in ordinary discourse. But how does his translating into the concrete (covertly) create this impression?

It does so, I believe, for the following reasons. First, since Moore does not say that the person is speaking with philosophical motives or intent, and since he does not, as Moore gives him to us, speak in the (less "concrete") phraseology we are accustomed to hear from Cartesian sceptics, there is nothing to suggest that the person is speaking as a philosopher. Second, the person we are invited to consider would seem (because of the lines that Moore gives him to speak) to be recommending or expressing an 'isolated' doubt, as we do in ordinary thought or discourse, i.e., a doubt *only* about Moore's hands or *only* about his state of dress, and not also a doubt about the whole of the external world. It therefore looks as though this person ought to have grounds that are appropriate for doubting precisely that Moore has hands (or is dressed), and such grounds would, of course, be ordinary (and not philosophical) grounds. Third, since the doubt (or seeming doubt) that we are asked to contemplate is apparently about whether Moore is dressed or naked (and in the other case, whether Moore has hands or has lost them through accident or amputation),²³ the suggestion is that the person we are being invited to consider is not to be thought of as being sceptical about having a body (or about

23. Moore does not make this explicit in the passage quoted above, but several paragraphs later he writes "If one of you suspected that one of my hands was artificial he might be said to get a proof of my proposition 'Here's one hand and here's another', by coming up and examining the suspected hand close up. . . ." "Proof of an External World," p. 149.

Moore's having a body). But this suggests that the person we are being invited to consider is not the Cartesian sceptic—or at any rate, that he is off duty as a philosopher and is engaged in ordinary discourse. And finally, since the person we are being invited to consider seems to have no doubt about having a body (and since Moore does not warn us against doing so) we are inclined to picture him as having no doubt that he is on the scene with a full complement of properly functioning bodily senses. (Recall here Moore's saying, in the passage from "Certainty," ". . . for me, now, in full possession of my senses. . . .") And this, too, leads us to think of him as being someone other than the sceptic, as someone who does *not*, as he speaks (dubiously) about the matter at hand, have in mind any philosophical grounds for doubt.

All of this, and perhaps more of the same, is suggested to us by Moore's translating into the concrete. This, then, explains why Moore found that by translating the sceptic's conclusion into the concrete he had robbed it of all its plausibility. For what he has in fact done is this: he has covertly scuttled the sceptic and the sceptic's conclusion and has given us (and himself) something quite different to contemplate, namely, the picture of someone with neither ordinary grounds for doubt nor philosophical grounds for scepticism who nevertheless goes about saying that (or speaking as though) he is in doubt or finds various things doubtful. This is, assuredly, an altogether ridiculous picture, but hardly a picture that should embarrass the sceptic.

To put the matter in another way, the 'expression of doubt' which Moore puts into the mouth of this imagined person cannot be *both* an instance of the sceptic's scepticism ("what it means in particular instances") and also something said in ordinary, non-philosophical discourse. And yet Moore needs to have it both ways. For were he to allow that it is *not* to be thought of as an instance of the sceptic's scepticism, then it would be irrelevant to rebutting such scepticism, whereas if it *is* to be thought of as an instance of the sceptic's scepticism (and therefore *not* something said in ordinary discourse), then it cannot be thought of as being (cannot be heard to be) absurd. Or conversely, insofar as it *is* to be thought of as being (or heard to be) absurd, ridiculous, irrational, we must think of

it as something said in ordinary discourse (since philosophers' talk in its context *doesn't sound that way*), but in that case it is not something said with philosophic intent and so is not an instance of the sceptic's scepticism. So Moore needs to have what he cannot have: he needs to have it both ways. The fact that he *thought* he had succeeded in confuting the sceptic can only lead us to conclude, then, that he was (unwittingly) equivocating in his thoughts between these two incompatible alternatives.

We now have, I think, the explanation of how Moore, despite holding the representative theory of perception, could have thought he knew things the sceptic insists one cannot know. Moore thought this because he began by thinking: "Either I know such things or (as the sceptic claims) I do not," and because he then thought, as a result of his translating into the concrete, that he had detected an absurdity in (what he takes to be) the sceptic's claim or conclusion. This, as we have just seen, was a mistake, but having made that mistake, Moore concluded that it was only reasonable that he should say: I know this is a finger, I know I have hands, I know I am dressed. But just as Moore must have been equivocating in his thoughts about what it would be ridiculous or absurd to say, so too he must have been equivocating in his thoughts about the nature of these knowledge claims. For if they are to be taken (as Moore would have us take them) as contradicting the sceptic, then Moore cannot hope to secure our agreement with him in this matter of what he knows by claiming that it would be ridiculous to contradict his knowledge claims, for if his knowledge claims do contradict the sceptic, then were we to contradict Moore's knowledge claims, we would be speaking with the sceptic, and that, as we have seen, would not be ridiculous: we would sound, not like lunatics, but like philosophers.

The source of all this confusion is, as we have seen, Moore's technique of translating into the concrete. This appears to be a harmless, and even a valuable, philosophical technique, and yet in the present case it has served only to mislead and confuse. This is so because it tempts us to ignore the difference between speaking as a philosopher and speaking as a plain man. In short, it takes advantage of our weakness for that old, but dubious, philosophical notion: the proposition.

Justification of Belief:

A Primer

CARL GINET

I

I assume that for any given person and time, all propositions can be divided exactly into two sets: those believed by that person at that time and those that are not believed by that person at that time. This is, of course, idealizing a bit, ignoring the fact that the concepts of a belief and of a proposition are vague, but it will do no harm for present purposes.

The concept of a belief and the concept of a proposition each deserves a treatise unto itself. Here I have to take them for granted, except for the following few remarks. What I mean by a proposition is something that is true or false and that has the same truth-value at all times. To say that a person believes the proposition that *p* is the same as to say that the person believes that *p*; and to say that is to imply that the person understands the proposition at least to some degree. A belief is a dispositional state that a person may be in even when not manifesting that state in any mental or bodily act. The term "belief" covers a wide range of phenomena. There are beliefs that the believer never explicitly articulates (and perhaps could not articulate), that are deeply embedded in the believer's culture or the human way of life, imbibed early in life (perhaps even given innately). Some beliefs (such as beliefs as to what one is currently perceiving) arise and pass without any conscious articulation of them or of reasons for holding them, but the subject could easily articulate them, along with reasons, if the occasion arose. And there are beliefs that a person has arrived at through a process of inquiry and deliberation. Belief can have varying

degrees of strength, from being inclined to think that . . . to being confident that . . . , but I shall here ignore the weaker degrees and use "belief" (and its cognates) as short for "confident belief" (and its cognates).

II

A belief can be justified or unjustified. It is the negative notion here that issues the orders. One *ought* not to believe a proposition if, and only if, one *lacks* justification for believing it. Given a particular proposition and person (whose beliefs can be appraised in this way), if it is not the case that the person ought not to believe the proposition, then she is justified in believing it, if she does so, and in any case, she has justification for believing it. This is so even if she cannot be said, in any ordinary sense, to have *grounds* or *reasons* for believing it, or to have *evidence* of its truth. All that is required is that it would be wrong to reproach her for believing the proposition.

Any belief that one cannot help having must be justified, in a sense. If one cannot help it then one cannot be properly reproached for it. This notion of justification is broader than epistemologists are used to. There is a useful narrower sense in which some beliefs that are irreproachable, because they cannot be helped, are nevertheless *not* justified. For example, as a result of a bad experience with a large dog, a man is for a time unable to resist believing that any large dog that approaches him is about to attack him. Or a hypnotist is able to cause a subject to believe that, despite appearances and what past experience might lead him to expect, every tree in the United States has been cut down overnight. Such beliefs, we want to say, are *not rationally justified*. It is *rational justification* that epistemologists have been mainly interested in and that will be the main concern of the rest of this paper. Hereafter, unless the context clearly indicates otherwise, "justified" and cognate terms should be taken in this narrower sense.

The fact that a belief is compelled does not mean that it is justified *only* in the broader sense. Many of our rationally justified beliefs are ones that, in the circumstances, we could not have helped acquiring and could not cast off at will—for example, my

current belief, as I look out the window, that the sun is shining. It is generally a useful test of whether or not a belief is rationally justified to ask whether or not the subject would be justified in having it even if she could help doing so. But it may not be infallible. The subject should be able to apply the test to her own undiscardable beliefs. But perhaps she cannot even imagine what it would be like not to have the belief in question and perhaps the cause of this itself has justificatory force. The proposition may be one of those that one cannot entertain without (as Descartes put it) clearly and distinctly perceiving it to be true; or the belief may be a linchpin for the subject's whole belief system so that if it were unjustified then so would be most of the framework of beliefs on which she continually relies. (Such a belief might be in certain very general regularities or a more specific belief as to how human life works; it might be innate or instinctive or inculcated early and be continually reinforced by the culture.)

Can we say that every person's beliefs at every time divide into those that are then (rationally) justified and those that are not? No. The problem is not just borderline cases; those we can ignore. The problem is that the application of either of the pair "justified"/"unjustified" to a particular person's beliefs presupposes that the person satisfies a certain condition that need not be satisfied by every person who has beliefs. The beliefs of a small child, for instance, cannot be said to be either justified or unjustified because the child altogether lacks any concept of the justification of belief, any sense of that sort of consideration in the formation and alteration of belief. (For similar reasons, small children cannot be said to be morally justified or unjustified in, deserving or not deserving of moral reproach for, any of their actions.) Unless a person has at least a *modest* hold on the notion of justification of belief, it would be absurd to reproach her for failing to measure up to standards of rational justification in some belief she has.

What we can say, I believe, is that for any person who is qualified to have her beliefs assessed in the dimension of justification, and for any time, that person's beliefs at that time divide into those that are then justified and those that are not (ignoring borderline cases).

Trivial as this assertion may seem, it is not uncontroversial. There have, for instance, been subjectivists who would deny that any proper sense can be made of saying of a belief that it is, or is *not*, justified. On their view, we may use terms to express our attitudes towards beliefs, but they cannot signify a property or status that belongs to a belief on the basis of objective criteria. I do not know how to show that such a view is wrong, except by developing an account of the objective criteria for the application of these terms that provides a generally satisfying explanation and illumination of their intuitive use. I cannot attempt that grand task here, so I will merely record my working conviction that the subjectivist view is wrong.

Another line of objection notes that for a great many of a qualified person's beliefs it will often be odd, or fail to make sense, to say that the belief is justified or that it is unjustified. For example, it will be odd for me, or anyone else, to say, out of the blue, that my current belief that I exist or that I have two hands is justified or that it is not justified. Having described such a situation as one in which it fails to make sense or is a misuse of language to say either thing, or one in which the question of which is true fails to arise, one may be tempted to infer that it is a situation in which there is no fact of the matter.¹ This does not strictly follow and it must be presumed mistaken if there is a better explanation of the oddity of the saying. It seems to me that there is. It is the same sort of oddity that there is in expressing, apropos of *nothing*, the indisputable truth that $1 + 1 = 2$ (or, for that matter, the indisputable falsehood that $1 + 1 = 3$ or the disputable falsehood that I am now neither justified nor not justified in believing that $1 + 1 = 2$). It is simply the oddity of saying that which the context

1. Norman Malcolm endorses an inference like this in regard to statements that I know something or that I do not know it, in his paper, "Moore and Wittgenstein on the Use of 'I know'" in *Essays on Wittgenstein in Honour of G. H. von Wright*, Jaakko Hintikka (ed.), *Acta Philosophica Fennica*, 28 (1976); reprinted in Norman Malcolm, *Thought and Knowledge* (Ithaca, New York: Cornell University Press, 1977).

There are, I am sure, other points in the present paper with which Malcolm would strenuously disagree. Well, he has only himself to thank for my being interested in such topics. I have him to thank for these and many others among my philosophical interests, and for most of my philosophical ideals.

gives no point to saying. Language is being misused in the sense that obvious principles of rational discourse are being violated. There is senselessness in the sense of pointlessness.

The other explanation of the senselessness—that there is no fact of the matter—has one particularly awkward feature (pointed out by H. P. Grice in his *William James Lectures*). On this explanation, whether or not the sentence in question expresses a true-or-false proposition is relative to the circumstances surrounding its utterance, rather than to the situation that the utterance purports to be about. It becomes possible that, relative to one set of possible circumstances for uttering “*s* is justified in believing that *p*”, there is a fact of the matter as to whether or not *s* is justified in believing that *p*, but, relative to another set, there is not, even though the facts about *s* that determine the fact of the matter in the first case are also there in the second case. For any person *s*, time *t*, and proposition *p*, it is no doubt possible to imagine circumstances that might arise in which there would be point in saying either that *s* was at *t* justified in believing that *p* or the contradictory. If that is so, then we have a way of defining what it is for it to be true absolutely that *s* is justified at *t* in believing that *p*. It is for the facts about *s* and *t* to be such that, should circumstances arise that would give point to saying it, they would guarantee the truth of what was said.

Finally, a difficulty for our trivial-seeming claim may be seen in the fact that we can never actually take on the whole of a person's beliefs and, starting from no assumptions at all as to which are justified, determine which are and which are not. This may be thought to show that the notion we have is not that of a belief's being categorically justified, but rather that of a belief's being justified *if* certain other beliefs of the subject, whose truth the subject has taken for granted, are justified. That conditional predicate, it may be thought, is really the one we always apply, and expresses the only notion of justification that we have.² Again, the view is not really supported by the con-

2. I find such a view suggested in some things Michael Williams says in his book, *Groundless Belief* (New Haven: Yale University Press, 1977). He asks, “What else could justification consist in if not bringing accepted beliefs [whose justification is not in question] forward in support of the propositions

sideration that suggests it. Notice that it is also true that we never start from scratch in assessing the truth-values of propositions. We never start by making no assumptions whatever as to the truth-values of any propositions and then proceeding to determine the values for some of them. We always take for granted the truth-values of a great many propositions and use these assumptions in investigating the truth-values of the ones we are currently interested in. But this is not a good reason for saying that we do not have a notion of a proposition's being true categorically, but only a notion of a proposition's being true if others are true.

Indeed, it is not easy to see how to make sense of the view that the predicates "justified" and "unjustified" do not apply to beliefs categorically but only conditionally. How can the conditional "s's belief that p is justified if s's belief that q is justified" have a truth-value if neither its antecedent nor its consequent can have one? How can establishing the conditional "I am justified in believing p if I am justified in believing q" by itself be pertinent to my deliberation as to whether or not to believe p? Surely I have nothing relevant to this decision until I can *detach* the consequent, in virtue of having established the antecedent (as well as the conditional). It would not be helpful to this sort of view to say that the conditional that is really being affirmed in all categorical-looking attributions of justification does not employ the notion of justification in its antecedent but just that of belief: "If S believes that q, then S has justification for believing that p". Here it is clear that the antecedent can be categorically true and so, therefore, can the consequent if the conditional can.

So let us take our claim, that all of a qualified person's beliefs at a given time divide into those that are then justified and those that are not then justified, to be as innocent as it looks.

to be justified?" and he doubts that there is any sense to the idea of "justifying the whole thing", which would seem to mean the idea of saying what determines which set of a person's beliefs is the totality of her justified beliefs (p. 99). I hesitate, however, to attribute any such view to him, since his book is not much interested in positive characterization of the concept of justification.

III

It can happen that at a particular time a person has justification for believing a certain proposition but does not then actually believe it. For at least some propositions, it is possible that at a time when one does not believe the proposition there obtains a condition such that if one were then to believe the proposition then one would, simply in virtue of that condition, be justified in that belief. For example, someone may read in the newspaper that the number on the ticket she owns is the winning number in the lottery, and in consequence have justification for believing that she has won the lottery; but, out of extreme fear of disappointment, she refuses to believe this until the money is actually handed over to her. Or a person who is not abnormally given to misremembering may occasionally be unreasonably distrustful of her memory: although it seems to her that she clearly remembers putting matches in her backpack, she hesitates to believe this until she has checked it. We can say that, for any given person and time, all propositions divide into those that the person then has justification for believing and those that the person does not then have justification for believing. Let us refer to the set of propositions that person s at time t has justification for believing as J_{st} .³ What we have just pointed out is that J_{st} need not be included in the set of propositions that s believes at t (which we can refer to as B_{st}).

I said that cases where a member of J_{st} is not also a member of B_{st} are ones in which there obtains at t a condition such that if the proposition in question had been a member of B_{st} , and the condition obtained, then s 's belief in it would, in virtue of that condition, have been justified. It might be thought that this description of such cases must be incomplete. It might be thought that if there is such a thing as having justification for believing what one does not actually believe, it will have to be

3. I am assuming that there is always just one such set, that there are not alternative maximal justified sets, each of which is such that s at t has justification for believing all of its members together and it is not included in any larger set of which the same is true. This assumption is convenient for economy of exposition but it is not, I think, essential for any of the points I wish to make. Cf. note 7 below.

described in the following way: there obtains a condition such that if *s* had been *caused* (prompted, led) to believe the proposition by that condition then *s* would have been justified in believing it. This contention would be motivated by the idea that the justification of a belief must always be a *matter of how the belief is caused*, must consist in the belief's being produced or sustained by the right sort of factors. But I am disinclined to accept this idea. It seems to me no more likely to be true than the corresponding idea about the moral justification of action. The considerations in terms of which it is explained why some action is morally justified—e.g., that it was justifiedly believed by the agent not to harm anyone or not to be in violation of the agreement—need have little to do with what actually motivated the agent to perform the action. It need not even be the case that the agent would not have performed the action had no such justifying factors been present. A similar situation may obtain, it seems to me, with respect to the justification of belief. Suppose that the lottery winner, though not prompted to believe that she had won the lottery by her knowledge of what she saw in the newspaper and of what was said in a telephone call to her from a lottery official, was subsequently prompted to believe it by her being told it by a gypsy fortune teller (in whom she has irrational confidence). It seems to me that she is justified in this belief. She is not justified by the gypsy fortune teller, but rather by those factors that, if she did not believe it, we would cite in explaining why she nevertheless has justification for believing it. She should not be reproached for *the belief*. After all, she could respond to such a reproach by citing the newspaper report and telephone call, her knowledge of which we must admit to have justifying force. She may, however, be reproached for being *moved* to believe by her knowledge of what the fortune teller said. So it does not seem to me always necessary that the condition constituting justification of a belief has to be something on which the belief is dependent—or, more accurately, has to have the form that the belief is dependent on such-and-such factors. Instead of any actual causal connection, what is always necessary, as I will argue below, is that the condition constituting justification be directly accessible to the subject; so that the subject *could take account of it, could be influenced by it*, if she wished

to ensure that her belief was justified and she recognized the relevant principles of justification.

The property of belonging to J_{st} (for some s and t) is a *supervenient* property of a proposition. That is, it is a property it has in virtue of other properties it has that can be specified without attributing *that* property to anything: it can acquire or lose this property only because of a change in some other thus neutrally specifiable property. So there must be correct principles of justification that can, for any person s (qualified to have justified or unjustified beliefs) and time t , take us from sufficient neutrally specified information to a determination of the membership of J_{st} . Let us call these the J-principles. The J-principles can be thought of as telling us, for any given proposition, what counts as a sufficient condition for that proposition's belonging to J_{st} (for some s and t) and what counts as a necessary condition. (If we know everything of that sort that the J-principles have to tell us, then we know everything they have to tell us.)⁴

4. I pointed out earlier that any person s for whom there is a set J_{st} must at t have at least a minimal grasp of the concept of justification of belief. This means that the person must believe at least some of the correct J-principles. (A person may believe J-principles without being able to articulate them. That a person believes a certain J-principle can show itself in the person's being sensitive in the appropriate way to the considerations referred to in the principle in deciding what to believe or in appraising the justification of her or others' beliefs.) This in turn means that any sufficient condition for membership in a J-set that the J-principles lay down will have to include this condition that the subject believes a minimal number of the J-principles. Does this mean that some J-principles will be involved in a kind of self-reference, referring to a totality of items of which they themselves are members? Even if it did, it is not clear that it would be objectionable, but it seems that such self-reference can be avoided. It could be arranged that all of the J-principles but one refer simply to people who are qualified to have justified or unjustified beliefs, rather than to people who accept a minimal number of the J-principles. Then the remaining J-principle could say that to be qualified to have justified or unjustified beliefs is to accept some minimal subset of those other J-principles. Thus the total set of J-principles would entail propositions of the form "If (or only if) condition C obtains and s believes a minimal number of principles P_1, \dots, P_n , then q belongs to J_{st} ", but none of these propositions would be among the principles P_1, \dots, P_n . Rather, among P_1, \dots, P_n there would, for each of those propositions, be a corresponding proposition of the form "If (or only if) condition C obtains, then q belongs to J_{st} ".

IV

It must be possible for there to be an epistemically perfect person. That is, it must be in principle possible for there to be an s and t such that s has no unjustified beliefs at t — B_{st} is included in J_{st} —and, moreover, as t changes s keeps B_{st} included in J_{st} by knowing the correct J -principles and having a sufficiently strong will to adhere to them (and no irrational compulsions or external interference to defeat her). It should be possible that knowledge of the J -principles would enable s continually to know the boundaries of J_{st} and, given sufficient will, *thereby* (for the right reasons) continually to keep B_{st} within those boundaries. If this were not so, then the concept of justification would scarcely be coherent: the principles of justification would lay down a guide to forming and altering belief that even a person with a complete grasp of them and commitment to them, and with the best will in the world, might be unable to follow correctly or unable to know that she was following correctly. Any account of the moral justification of *action* that had the corresponding consequence would, clearly, have to be ruled unacceptable or else taken as showing that the notion of moral justification itself is incoherent and unacceptable. I see no reason why we should not say the same for the notion of justification of belief.

I wish to note two important consequences that follow from the point that such epistemic perfection is possible. First, for the propositions belonging to J_{st} (for any s and t), there must be neutral facts sufficient to make them members of J_{st} that are *directly accessible* to s at t . That is, each such fact must be such that s needs at t only to give clear-headed attention to the question of whether or not a fact of the relevant sort obtains in order to be aware that it does. They cannot be such that they could obtain while s was unable to know that they did without a process of investigation. For if any such facts relevant to determining the membership of J_{st} were thus beyond s 's direct grasp at t , then it could be that, no matter how thorough s 's grasp of all the correct J -principles and no matter how strong s 's will to adhere to them, it is beyond s 's power to ensure that $B_{st} \cap \overline{J_{st}}$ does not expand as t changes. For then s 's position at some time might be that s could not then and there determine for some propositions

whether or not they belonged to J_{st} . Their status could change, from being in to being out of J_{st} , or vice versa, without s 's being able to detect the change when it occurs. A set of instructions aiming to give signs by following which one will stay on the right path must use signs that can be detected from the relevant points on the path.

A J-principle specifies a condition relevant to determining whether or not a proposition belongs to J_{st} . Let us say of one that does this in terms of neutral facts directly accessible to s at t that it is a *usable* J-principle. Then the first important consequence of the point made in the paragraph before last can be put this way: some set of correct J-principles that is complete, in the sense that it can determine J_{st} for any s and t , must all be usable.

The other important consequence of the possibility of epistemic perfection is that some complete set of usable and correct J-principles is such that anyone who accepts them is justified in doing so. If epistemic perfection is possible, then it is possible for the beliefs of someone who accepts such a set of J-principles and follows them perfectly *all* to be justified. This includes the beliefs in those principles. But if belief in those principles is justified in that case, it is hard to see how it could fail to be justified in all other cases, where the believer accepts the principles but does not adhere perfectly to them (through inability or reproachable failure) or has beliefs that are in accord with the principles only accidentally. That a person fails to adhere faithfully to a principle she believes, or is in accordance with it unintentionally, is certainly not sufficient to make it the case that her belief in it is not justified. So S 's acceptance of those principles is justified if she does, and also if she does not, always follow them faithfully; that is, it is justified in any case. So the qualifications for membership in J_{st} for *some* complete set of usable, correct J-principles are quite minimal: s has at t only to believe the principles (or, even less, s has only to understand them, but must do at least that; see below, Section VII).

It does not follow that this minimal condition suffices to make *every* correct and usable J-principle belong to J_{st} . There might be such a principle that is contingent. The principle gives correct results in just those possible worlds, of which the actual

world happens to be one, in which there obtains some general (neutrally specified) fact G . Perhaps some J-principles that prescribe situations in which a proposition is justified by non-deductive inference are of this sort. A justification for accepting any such contingent J-principle, P , must be one of two sorts. It could be by (deductive) inference from P if G , and G , where one has independent justification for G . Or it could be that one cannot help accepting P and that one's disposition to be guided by P in forming or altering one's beliefs is inborn, or trained into one at an early age, in such a way that one could simply never become able to suspend it while waiting for independent justification for G . (But one could, in this case, have justification for G that is dependent on P , that is by inference from G if P , and P .) Either way, one's justification does not consist merely in one's understanding P , or even one's believing P . If this is right, then what does follow from the conclusion of the preceding paragraph is this: there must be a complete set of usable, correct J-principles that are all non-contingent.⁵

V

In the rest of this paper, I want to take up a question that has been brought to the fore by the debate in recent epistemological literature over "foundationalism" vs. "coherentism". To what extent is the membership of a J_{st} determined by internal relations among propositions?

An internal relation among propositions is one that follows from the content or forms of those propositions. For example, among the propositions:

5. The weaker thesis, that there must be some complete set of correct J-principles (usable or not) that are all non-contingent, can be shown by an independent and more obvious argument. If J-principle P gives correct results in worlds in which G but not in worlds in which *not-G*, then P iff G is itself a J-principle that gives correct results no matter how the world is, a necessary J-principle. The set of all necessary J-principles derivable in this fashion from contingent ones, plus any other necessary J-principles there may be, yields all the same results for the actual world as does any set of J-principles that holds in the actual world. Thus, if there is a set of J-principles sufficient to determine the membership of any J_{st} in the actual world, there must be such a set all of whose principles are non-contingent.

- (1) Every Persian rug is rectangular.
- (2) My only Persian rug is rectangular.
- (3) My only Persian rug is circular.
- (4) If my only Persian rug is rectangular, then it is not circular.
- (5) Every Persian rug I've observed is rectangular.

there obtain (among others) the following internal relations:

- (i) (2) is a universal instantiation of (1).
- (ii) (5) is deducible from (1).
- (iii) (2) is the antecedent of (4).
- (iv) The negation of (3) follows by *modus ponens* from (2) and (4).
- (v) (1) and (5) are both instances of the propositional form 'Every F is G'.
- (vi) (1)–(5) are all about Persian rugs.

And we may allow that a single proposition can enter into a "monadic internal relation" in virtue of having a certain sort of content or form. So further examples of internal relations among (1)–(5) would be:

- (vii) (5) is an instance of the propositional form 'Every F I've observed is a G'.
- (viii) (4) is a conditional.
- (ix) (1) is about Persian rugs.

Some of the internal relations among members of a J_{st} may figure in their qualifications for membership. For instance, there may be *sufficient* conditions for membership that have the following form: " p belongs to J_{st} if p has (internal) relation R to other members of J_{st} ". Let us call any such principle an *inference* principle and the relation referred to in its antecedent, an *inferential* relation. More specific forms of inference principles might, for example, include the following: " p belongs to J_{st} if the conjunction of q and p is deducible from q belongs to J_{st} " or " p belongs to J_{st} if the conjunction of q and p has the relation R (not implying deducibility) to q belongs to J_{st} and there are no members of J_{st} to which *not- p* has the relation T ". It is

plausible to suppose that justification by non-deductive inference would be subsumed under principles of this latter form.

The J-principles may entail *necessary* conditions for membership in J_{st} of the following form: " p belongs to J_{st} only if p has (internal) relation R to the other members of J_{st} ". For example, it is plausible to suppose that any given proposition is a member of J_{st} only if no other member of J_{st} is the negation of that proposition. Let us call any J-principle of this form a *coherence* principle and the internal relation referred to in its consequent, a *coherence relation*.

(Note that, since a sufficient condition must entail every necessary condition, every inferential relation must be defined so that a proposition's having it to members of J_{st} guarantees that the proposition has all the coherence relations to members of J_{st} . If we start with a *coherent* set and enlarge it by application of inference principles, the resulting set will be coherent.)

Let us say that J_{st} has a *foundational structure* just in case it divides into two non-empty subsets, the *foundational* and *non-foundational*, such that the justification for every member of the non-foundational subset is, in some sense or other, founded on or derived from the justification of some members of the foundational subset, but none of the members of the foundational subset derives its justification from members of the non-foundational subset. If one specifies the sense of "founded on or derived from" as "attaches to it only because of an inferential relation, or chain of inferential relations, it has to", then, it seems to me, one has specified a type of foundational structure that the typical J_{st} is very likely to have. For it requires only that the J-principles include some non-redundant inference principles—non-redundant in the sense that if they are applied to the largest subset of J_{st} yielded solely by other, non-inference J-principles (together with the neutral facts), then the inference principles will generate new members for J_{st} . In that case, the maximal subset of J_{st} generated by the non-inference principles is *foundational* in the sense specified and the additional members of J_{st} generated by application of the inference principles to the foundational set comprise the non-foundational subset: every member of the latter is a member of J_{st} only because of an *inferential* relation it has to members of the foundational set,

but of no member of the foundational set is it true that it is a member of J_{st} *only* because of an inferential relation it has to members of the non-foundational set. A member of the foundational set *may have* an inferential relation to members of the non-foundational set or to some combination of these—even to a set that contains itself—but this will not make it non-foundational, because its membership in J_{st} is guaranteed by criteria independent of that inferential relation.

Recall the point made earlier (in Section IV) that some complete set of usable and correct J-principles is such that it is included in J_{st} if s believes (or even just understands) its members. This means that such a set of J-principles will, if s believes its members, belong to the foundational part of J_{st} . For their justification will not depend on any inferential relations they have to other members of J_{st} .

VI

"Foundationalism" in epistemology has come to mean a sort of view that is stronger than the view that a J_{st} (typically) has a foundational structure (in the sense just defined). It connotes also a claim of some further special status for the foundational members of J_{st} . Common to these further claims is the idea that the foundational members are determined at least partly by *external* considerations—ones independent of internal relations of these propositions—having to do with what is true of s and t (Descartes, for example, would cite what s clearly and distinctly perceives at t). Attacks on foundationalism seem to focus as much on the further claims as on the claim of foundational structure. And it is interesting that the label "coherentism", which has become common for an anti-foundationalist position, suggests an emphasis on internal relations as determinants of J_{st} and suggests that such determinants can do at least some of the job that a foundationalist would have done by external considerations.

One extreme to which foundationalism might, conceivably, be carried is that of holding that the foundational members of J_{st} are determined entirely by external considerations and need satisfy no coherence requirements whatsoever: for each founda-

tional member of J_{st} there is a condition sufficient for its belonging to J_{st} that is independent of any of its internal properties or internal relations to other members of J_{st} . One extreme to which coherentism might be carried would be that of holding that the only sort of factor that needs to be taken into account in determining J_{st} is that of internal properties and relations of propositions. I think it beyond doubt that the truth must lie somewhere between these two extremes. In the remainder of this paper, I would like to explain why and to try to narrow a bit further the area where the truth must be.

The extreme foundationalist view cannot be right if there are any coherence principles among the correct J-principles, that is, any internal relations that any member of a J_{st} must have to the other members. And surely there are: for instance, the relation (mentioned above) of not being the contradictory of any other member of J_{st} . It is unlikely that such an extreme view has been held. We do *not* have such a view in any foundationalist view that holds that all *truths* of a certain sort belong to the foundational subset. (One variant of this would be the view that the membership of the foundational subset of J_{st} comprises every proposition p such that at t s believes that p and it is necessarily true that if at t s believes that p then p .) For in the requirement that they all be truths the view secures the satisfaction of at least some legitimate coherence requirements. Nor would we have this extreme in a foundationalist view that held that all the propositions believed by s at t having a certain sort of content are members of the foundational part of J_{st} . For this sufficient condition is specified partly in terms of an internal relation among the members of the subset: their all having a certain sort of content.

Consider now the extreme "coherentist" view, that if you know enough about the internal features of propositions (and also the correct J-principles) then you have all you need to determine J_{st} for any s and t .⁶ On such a view—and indeed on any

6. Lawrence Bonjour describes the view he presents in "The Coherence Theory of Empirical Knowledge," *Philosophical Studies*, 30 (1976):281-312, as claiming that "the epistemic justification attaching to an empirical proposition always derives *entirely* from considerations of coherence" (p. 308, n. 3, emphasis mine). This seems to place his view at what I have called the co-

view—the J-principles will have to include other than inference and coherence principles. If you consult the forms in terms of which I defined inference and coherence principles, in the preceding section, you will see that from such principles alone, together with the facts about internal properties and relations of propositions, nothing can follow as to what propositions belong to any J_{st} . But we can construct a set of J-principles formally capable of generating conclusions of the form 'p is a member of J_{st} ' without using principles that consider any sorts of neutral facts other than internal relations among propositions. For instance, besides the inference and coherence principles, we could have a principle of the following form: let A be the set of all and only those sets of propositions (or all and only those sets drawn from some totality of propositions specified by their internal properties) each of which satisfies all the coherence and inference principles; then any maximal member of A —i.e., any member of A not included in any other member—is a (maximal) J_{st} .

It is clear, however, that no view can be correct that says that J_{st} can be determined by applying to the set of *all* propositions a function defined entirely in terms of internal relations of propositions. Any such view has the absurd consequence that J_{st} is exactly the same set for every s and t , that one could, in principle, determine what any person was justified in believing at any given time without knowing the person or time in question or, indeed, any contingent facts about the world at all. The J-principles cannot be like that. If anything in this area is obvious, it is that what I have justification for believing now is different from what I had justification for believing at this time yesterday and from what Leibniz had justification for believing at noon on January 1, 1700 (Hanover time).

There can be no objection to thinking of the J-principles as determining J_{st} by applying a function defined entirely in terms of internal relations among propositions to *some* set of proposi-

herentist extreme. But the details of the view he actually presents do not seem to justify this description, and he himself, in a sentence adjacent to the one from which I just quoted, says that his view "does not hold that the only factor which determines the acceptability of a set of empirical propositions as putative empirical knowledge is its internal coherence."

tions. But they will have to specify this "feeder" set for the function, not merely on the basis of internal considerations, but also on the basis of what is contingently true of s and t .

VII

Let us in fact think of the J -principles as saying that $J_{st} = D(F_{st})$, where F_{st} is the "feeder" set and D is a function defined entirely in terms of internal relations among propositions. Clearly, D will have these properties if $D(F_{st}) = C(F_{st}) \cup I[C(F_{st})]$, where " $C(x)$ " denotes the maximal subset of x that satisfies the coherence principles,⁷ and " $I(x)$ " denotes the maximal set that can be generated from x by iterated application of the inference principles. Whether or not J_{st} has a foundational structure will then be a question of whether or not there are any non-redundant inference principles and, if there are, whether or not $C(F_{st})$ furnishes any grist for them. What kinds of external considerations play a role in determining J_{st} will be a matter of how the J -principles define F_{st} . (The notion of the "feeder" set F_{st} should not be confused with the notion of a foundational subset of J_{st} defined earlier. That $J_{st} = C(F_{st}) \cup I[C(F_{st})]$ obviously does not guarantee that J_{st} will include F_{st} .)

We have seen that F_{st} must be defined at least partly in terms of external considerations. What sorts of external considerations? A rather simple answer to this might tempt us. It might be thought that the only external factor we need to consider is the question of what S believes at t and, therefore, that F_{st} can be neatly defined as the set of just those propositions that s believes at t —i.e., B_{st} .⁸ (This defines F_{st} partly in terms of which proposi-

7. Actually, nothing shown here warrants the assumption that there will always be no more than one maximal subset of F_{st} that satisfies all the coherence principles. In a situation that failed to satisfy this assumption there might be more than one way of dividing all propositions into those that belong to a (maximal) J_{st} and those that do not. It is not obvious *a priori* that such a possibility can be ruled out, but, as I said earlier (note 3), it is convenient, and (I believe) harmless for my purposes here, to ignore it.

8. R. Firth describes this position when he speaks of a coherence theory that limits "the class of basic warrant-conferring statements" for s at t to B_{st} . He suggests that this is the most plausible delimitation of this class. See his "Coherence, Certainty, and Epistemic Priority," *The Journal of Philosophy*, 61 (1964): p. 556.

tions of the form "At t s believes that p " are true and partly in terms of the internal relation that " p " has to such a proposition; an equivalent alternative would be to make F_{st} the set of all truths of that form and put the internal relation into the function D .)⁹

This suggestion is appealingly simple, but it will not work. The correct formula cannot be $J_{st} = D(B_{st})$, no matter how D is defined. If it were, we would have to accept that s can keep J_{st} the same over a stretch of time just by keeping her beliefs the same, no matter what other sorts of changes there may be in the meantime. But we should not accept this. Suppose I am watching a bird sitting on a branch, believing that I see such a thing. Suppose the bird flies away and my visual experience undergoes a corresponding change: it is as if I were seeing a bird fly away from the branch I saw it sitting on. But suppose I manage to keep my beliefs from changing accordingly: I continue to believe exactly what I believed a few moments ago when I saw the bird on the branch. I add no new propositions to the stock of those I believe, including none about what has happened to that bird since a few moments ago. (Never mind how I manage to do this or even whether or not it is in my power to do it; it needs only to be logically possible.) If the set of propositions I have justification for believing were at all times simply an internal function of what I believe, then that set would not change over the interval described. But surely it does change. Given the change in my visual experience (and the absence of any other unusual sense experience or memory phenomena) there is added to it the proposition that I saw the bird I was watching fly away; this is so whether I believe this proposition or perversely manage to keep from believing it.

The view under consideration would also mean that if I

9. Keith Lehrer, in his book *Knowledge* (Oxford: Oxford University Press, 1974), presents a view of essentially this kind. On his view (if I understand it correctly), $J_{st} = D(F_{st})$, where D is a function defined entirely in terms of internal relations and F_{st} is the set of all those propositions that describe s 's "corrected doxastic system" at t , that is, all those propositions of the form " s at t believes that p " that are true and would still be true if s were an "impartial and disinterested truth-seeker" (see pp. 198-208). The last clause means merely that the belief in question has not arisen out of such a motive as s 's very much wanting it to be the case that p .

could manage to change *only my beliefs* in a massive and coherent way then that would suffice to change the propositions I have justification for believing in an equally massive way. Suppose that nearly all of what I believe about my past is justified for me now. Suppose that I then suddenly shift from believing in that past to believing in one that is thoroughly different from it but equally coherent in itself and with my current experience, and suppose that I manage to do this despite there being no change in my memory impressions. They still deliver the same past I used to believe in but I now regard them as simply delusory: I think that I did not actually experience any of what I seem to remember experiencing. Imagine that my massive change in my beliefs and massive dismissal of my memory impressions is not related to any unusual developments in my perceptual experience or my memory impressions: I manage the feat by sheer will alone. The only unusual change is in the beliefs; everything else goes along normally. The bizarreness of this case may make it difficult for our intuitions to get a grip on it. One thing we would clearly *not* want to say about it, however, is that I could in this way, through unusual (and perhaps humanly impossible) control of my beliefs, massively and suddenly change what I am justified in believing. We might want to say that I have suddenly gone so crazy that "justified" and "unjustified" no longer apply to my beliefs, but we should not say that my new set of beliefs is mostly justified, just like my old set. Yet this is what we would have to say if we held that, for some internal function D , J_{st} is always $D(B_{st})$.

Clearly, other sorts of facts will have to be among those used as a basis for determining what belongs to F_{st} . The examples just considered suggest two other sorts that will have to be included. One comprises the facts as to what s 's subjective experience at t is as if she were perceiving (seeing, hearing, etc.); these we may call the facts as to what propositions of the form "I perceive X " are delivered by s 's perceptual experience at t . The other comprises the facts as to what it seems to s that she remembers having experienced or having come to know (e.g., it seems to s that she remembers having recently seen her glasses on her dresser, or it seems to s that she remembers having learned that any two sides of a triangle are proportional to the sines of

the opposite angles). I mean here that sense of "it seems to *s* that she remembers" that is implied by "*s* remembers" but such that "it seems to *s* that she remembers that *p*" is compatible with "*s* knows that not-*p*". Facts of this second sort we may speak of as the facts as to what propositions are delivered by *s*'s memory-impressions at *t*. The appeal of the idea that the feeder set F_{st} could be simply B_{st} may depend, in part, on the fact that in the actual world B_{st} typically includes a great many, if not all, of those propositions delivered by *s*'s sense-experiences and memory impressions at *t*. But this is only a contingent connection, and it would not be plausible to say that the *J*-principles apply only in those possible worlds where that connection obtains. The fact that a certain putative principle gives correct results for certain possible situations but not for others demands explanation in terms of more basic, less restricted principles.

Another sort of fact that the specification of F_{st} should take into account is the facts as to what propositions *s* *understands* at *t*. A person can have justification for believing a proposition only if that person understands the proposition well enough to be capable of believing it. *s*'s having justification for believing a proposition that *s* does not actually believe should mean that all *s* has to do in order to have a justified belief in it is simply to believe it. But if *s* does not understand it well enough then that is not all that *s* has to do: *s* must first acquire an adequate understanding of it. We said that J_{st} should be defined in such a way that, given that B_{st} is already included in J_{st} , all *s* needs in order to keep it so as *t* changes is a sufficiently strong will to do so (and a knowledge of the correct *J*-principles). These things should also be sufficient to enable *s* to keep J_{st} included in B_{st} (given that it once is).

If *s* already understands a certain proposition and has justification for believing it but does not yet believe it, then *s* can, normally, simply decide to believe it and forthwith do so. But one cannot acquire understanding of a proposition one does not yet understand by simply deciding to understand it and forthwith doing so. Justification is a normative notion and one should always be able to apply the consideration of justification directly to the decision whether to believe or not. If one ever had justification for believing what one did not understand, this considera-

tion could in that case have no such role in influencing such a decision, for the lack of understanding would prevent the question from arising. So J_{st} should be a subset of U_{st} , the set of propositions that s understands at t .¹⁰

This result can be secured if and only if the J -principles require F_{st} to be a subset of U_{st} and require the function D to be closed on U_{st} . Given our supposition about D , the internal relations it might use to enlarge J_{st} beyond F_{st} are limited to inferential relations. So inferential relations must be closed on U_{st} .

Requiring F_{st} to be included in U_{st} has a bearing on the way in which the facts as to s 's sense-experience and memory-impressions at t should be taken into account in specifying F_{st} . It should not be by saying that F_{st} includes the intersection of U_{st} with the set of propositions *reporting* s 's sense-experience and memory-impressions at t (" s 's visual experience at t was as if she were seeing X ", "It seemed to s at t that she remembered that p "). It may be, for many an s and t , that few or none of those propositions are contained in U_{st} . s may not have acquired the somewhat technical and sophisticated concepts of *sense-experience* and *memory-impression*. But this should not prevent s from being justified in her perceptual and memory beliefs. So what we should say is that F_{st} includes the intersection of U_{st} with the set of propositions *delivered* by s 's sense-experience and memory-impressions at t ("I see X ", "I remember that p ").

A great many members of this intersection will also be members of B_{st} . What other members of B_{st} should F_{st} contain? The complete answer to this may be difficult to discover, but at least part of the answer (to which I have already alluded) seems to me to be as follows. F_{st} should include those members of B_{st} mentioned earlier, where s finds herself incapable of not believing them and the reason for this incapacity has justificatory force: s "clearly and distinctly perceives" the proposition in question

10. This is not to say that propositions that do not belong to J_{st} but would do so if only they belonged to U_{st} are of no special epistemological interest as compared with those that would not belong to J_{st} even if they belonged to U_{st} . Moreover, I am inclined to suppose that there is a special class of propositions distinguished by the fact that their merely belonging to U_{st} is sufficient to make them members of J_{st} , and that this distinction is the basis of the distinction between a *a priori* and a *a posteriori* justification of belief.

or the belief is basic in s 's belief system. If the contents of such beliefs are members of F_{st} then they will be members of J_{st} —i.e., s 's beliefs in them will be justified—if and only if they are also members of $C(F_{st})$, the maximal coherent part of F_{st} . This is as it should be.

It is not the case, however, that all members of B_{st} should *ipso facto* be members of F_{st} . Suppose I were now to choose to believe the proposition that intelligent life exists on exactly fifty-seven planets in our galaxy, or were suddenly to find myself unable to resist believing it. If this proposition were granted membership in $F_{me,now}$ then it would surely also belong to $C(F_{me,now})$. It is hard to see what coherence principle could exclude it. Hence, given our supposition as to how J_{st} is a function of F_{st} , it would also belong to $J_{me,now}$. Yet in my present circumstances I would certainly not be justified in choosing to adopt this belief (and if I were unable to help believing it, this would be an irrational compulsion). Therefore, given our stipulation, it cannot be necessarily true that B_{st} is included in F_{st} .

VIII

The foregoing points fall very short of giving a complete and informative specification of F_{st} . But it has not been my ambition here to work out the content of the J-principles, either the coherence or the inference principles, or those determining the membership of F_{st} . I have tried only to lay down certain basic, general points about these principles.

If there is such a thing as the objective justification of belief, there must be necessary principles that determine, on the basis of neutrally specified facts that are directly accessible to s at t , the set of propositions that s has (rational) justification for believing at t —the set J_{st} . Belief in these J-principles (or something equivalent in the actual world) is always justified (so anyone who understands them has justification for believing them) and belief in at least a modicum of them must be an attribute of any person whose beliefs can be called either "justified" or "unjustified". The J-principles can be thought of as determining J_{st} in two stages. First, they specify, on the basis of facts about s and t , a "feeder" set of propositions, F_{st} . Second, they

provide a function, defined entirely in terms of internal relations of propositions, that takes F_{st} into J_{st} . This function has the form: $C(F_{st}) \cup I[C(F_{st})]$, where " $C(x)$ " is defined in terms of coherence relations and " $I(x)$ " in terms of inferential relations. J_{st} will have a foundational structure if there are non-redundant inference principles applicable to $C(F_{st})$. F_{st} must be included in the set of propositions s understands at t and the inferential relations must be closed on this set. F_{st} must contain the propositions delivered by s 's sense-experience and memory-impressions at t , as well as some propositions that s cannot help believing; but it need not contain all the propositions believed by s at t .¹¹

11. I am grateful to Sydney Shoemaker for a number of very helpful comments on an earlier version of this paper.

On Causal Knowledge

GEORG HENRIK VON WRIGHT

I

Let us first consider, in passing, an example of knowledge "based on induction", which is not causal. The example shall be that ravens are black. What does a man intimate (imply) about himself, if he says that he *knows* that ravens are black?

I think most of us, educated men, would say that we know this. Would we also say we know that *all* ravens are black? I think we should feel hesitant to stress the "all". This, I think, is significant. Saying that we know that ravens are black is not to say, by implication, that we know that there will never, never be an exception to a certain "uniformity of nature".

Why would we say we know that ravens are black? How many ravens have we seen? Most of us very few, if any. We have seen pictures of ravens; we have read about ravens in zoology books; we are familiar with what may be called the "proverbial" blackness of ravens. This is "secondhand" knowledge. At the basis of it is, of course, long familiarity with a species of birds the members of which invariably (or nearly so) have been found to be black. A member of the species would normally be identified on the basis of a few characteristics, of which blackness is one. "Are you sure the bird you saw was a raven?" "Yes, it was quite black, this big, and sitting on a carcass." If raven-like but not black birds are observed, we might lay these cases aside as "exceptions". Perhaps a zoologist or an experienced man in the woods could explain them to us. Single cases of this kind would not affect man's common knowledge that ravens are black. If

there occurred markedly many of them in, say, a hitherto little explored region of the world, we might have found a new species. Why could there not exist white ravens, since we know there are black swans?

Knowledge that ravens are black is part of our common, inherited knowledge. Knowledge that *all* ravens are black is not.

If the color is one of the characteristics by which we identify birds as ravens, does this not mean that blackness is logically connected with ravenness? And, if so, then surely *all* ravens are black.

But who would insist that blackness is a defining characteristic of ravens (or of ravens in such and such parts of the world)? At most a philosopher, who wishes to maintain that the reason why we *know* that ravens are black is that blackness is conceptually tied to ravenness. He would be wrong, however. An ornithologist, I think, would not insist on a conceptual tie here.

For some purposes, however, blackness could be *made* a defining characteristic of ravens. This could be some practical, transient purpose—such as counting the number of live ravens in a district. Or it could be some scientific purpose—such as creating a taxonomy. (But even given such purposes as those mentioned one would, presumably, be willing to admit “exceptions”.)

So, on what does our knowledge that ravens are black rest? Basically, of course, on extensive experiential data about the color of (members of) a certain species of bird. But also on such facts as the following: (a) The absence of specific reasons for doubting the universal truth of the blackness of ravens. We have, for example, no reason to think that in a certain unexplored region there exist non-black ravens. (b) The fact that we have some idea of how to cope with apparent counterinstances. Non-black ravens might be albinos, or they might belong to a different species from the ravens with which we have been familiar.

II

Let us now consider causal knowledge. A primitive example of causal knowledge is that if I put my hand in the fire, it will

hurt. Or that water in a kettle will start boiling if heated to a certain temperature.

These are things we *know*. But are they not too "primitive" and also too vague to be of much interest? What if for the second item of knowledge we substituted that pure alcohol boils when heated to 80° C? Or that water boils when heated to 100° C under normal atmospheric pressure (but not on the top of Mount Everest)? If the substituted items are said to be known, the question becomes relevant: Known to *whom*? To most people such items are only secondhand knowledge. (Also to most scientists.) But practically everyone of us has firsthand knowledge of the effects of heating a kettle of water on the stove.

Should we say that *what* everyone of us knows is that water under normal pressure boils at 100° C, although not too many of us know that *this* is what he knows, i.e. that this is the "exact" expression of the content of his knowledge?

This would not be right. The "common knowledge" that water boils, if heated, is not confined to situations when normal pressure obtains. It is therefore not "implicitly" knowledge that water boils if heated to a certain temperature, either. But neither is it unrestrictedly knowledge that "water boils when heated". It is common experience that if the flame under the kettle is weak and—as a person with rudimentary knowledge of physics would say—a loss of heat to the surroundings from the kettle and the water takes place, then even prolonged heating may not result in boiling. So, the efficiency of the heating must not be "offset" by prevailing circumstances and ongoing processes.

What then is it that we *know* about heating water and making it boil? In the individual case, we are absolutely certain that the water in the kettle which I placed on the stove will start boiling in a few minutes' time. This is what water under such circumstances does when there is fire under the kettle. We *know* this.

We know that, under certain circumstances, water boils when heated. We could not describe these circumstances in detail, but in the individual case we can normally tell with certainty whether they obtain.

I think this is how we should describe the epistemic situa-

tion. It seems that there is thus a "double knowledge" involved. There is knowledge of a generality, a "uniformity of nature". And there is a certainty, in the individual case, that the circumstances are such that this uniformity will manifest itself.

Why is it that it seems more appropriate to call our grasp of the concrete situation "certainty" rather than "knowledge"? This is not an idle question. Let us ask: What is it that we know about the individual situation which makes us sure that, if we light the stove and put a kettle of water on it, the water will within a couple of minutes start boiling? It seems that *nothing* in particular which we know about the situation is of relevance to this certainty. Relevant is rather the fact that we do *not* know anything about the situation which would make us think that, maybe, heating will not now be efficient in making the water boil. Our confidence in the working of the causal law on this particular occasion rests on the *absence* of reasons for thinking the contrary. We *know* that the law has worked on countless occasions in the past; we also know of occasions when the law did not work and have at least a rough idea how to characterize them; we have no reason to think that *this* occasion is "exceptional" rather than "normal"; *therefore* we are certain that the law will work here.

III

The physicist's knowledge that, at normal pressure, water boils at 100° C is not causal knowledge in the first instance, but logical knowledge based upon a convention fixing the centigrade scale. But "behind" this convention there is substantive knowledge about natural regularities—and at the very bottom there is our primitive causal knowledge that water can be made to boil by heating it.

Consider, however, some other liquid, the boiling-point of which is not by convention connected with a degree on the thermometer, but which is genuinely "measured". *Spiritus fortis*, at normal pressure, boils at approximately 80° C. I know this from books on chemistry. How do chemists know it? Most of them from books, I presume. But some have made experiments. Per-

haps this was in the course of their training. Then the experiments were not undertaken for the sake of checking or confirming the law—but rather for the sake of teaching the student experimental techniques. Deviant results would have shown that the circumstances had not been kept under the required control—not that the boiling-point of the liquid was, after all, *not* what the books say. But, of course, at the very basis of the knowledge which the chemistry books transmit, there are carefully conducted experiments undertaken in order to find out the boiling-point. I have no idea how many such experiments have been made. Perhaps in the case of some liquids, one was sufficient.

“Under” the chemists’ and physicists’ knowledge about the boiling- and melting-points of various stuffs, there is a mass of prescientific knowledge to the effect that each stuff changes its state of aggregation under roughly similar conditions of temperature—and not now at one and on another, seemingly similar occasion at a widely different temperature. And “surrounding” and “supporting” this knowledge is a body of scientific knowledge (about molecular compounds, atomic structure, etc.) which makes us expect, and partly explains, these facts about changes in states of aggregation. The fact that this supporting body consists of chemical and physical *theories* does not make the term “knowledge” inapplicable here.

We also have an idea as to when a claim to scientific knowledge can be *questioned*. New experimental techniques, for example, may enable us to determine boiling- and melting-points with still greater exactitude and thus to correct previous values. Experiments and observations under new conditions—say, extremely high or low pressures—may make us better aware of the restrictions to which observed regularities of nature are subject.

Questioning assumed scientific knowledge normally leads to “*improvements*” in our knowledge and not to “*overthrow*” of previous beliefs. This too we know. And this gives us a certainty that many of our present claims to scientific knowledge will never have to be completely renounced, but will at most become restricted relative to a bulk of old knowledge and old scientific techniques.

IV

Knowledge about boiling- and melting-points, whether scientific or prescientific, has the following characteristics which may be regarded as typical of causal knowledge: First, it is knowledge of relationships between changes in nature, e.g. that a change in temperature will cause a change in state of aggregation. Second, this knowledge is hypothetical in the sense that it pertains to what will happen, *if* something else happens. Third, it is relative to a frame of circumstances on the prevailing of which we can normally be certain in situations in which the causal relation is expected to hold, or its validity is put to a test.

Knowledge such as, say, that ravens are black or, generally, about typical features of members of a species and other "natural kinds" is different. It is not knowledge of how changes are related, but of how states are correlated. It is thus in a characteristic sense *static* as distinct from causal knowledge which is *dynamic*. Further, it is *categorical* and not *hypothetical*. We know that there are ravens and that they are black. If ravens become extinct this knowledge becomes "historical". We should then know that there was a bird, the raven, of which a black coloring was characteristic. If ravens were to change color in future, we should know that ravens used to be black. The fact that one can say truly "*if this bird is a raven, it is black*" does not make knowledge that ravens are black hypothetical. Ravens *are* black. So, *if* the bird you saw in the wood or which was brought here for examination was not black, it probably wasn't a raven. This illustrates one way in which a hypothetical can be "hooked on" to our general knowledge about the color of ravens.

V

The test of a causal uniformity requires that the circumstances under which it is supposed to hold, are, somehow, within our control. For example: if we want to test that liquid X boils at Y degrees under normal pressure, we must know how to test the pressure and preferably also how to regulate it and keep it

constant over the period of an experiment. We must also have some control over the loss of heat from the liquid to the instrument of measurement—know whether it has to be taken into account or whether it is negligible. Controlling the circumstances thus presupposes a great deal of “background” causal knowledge and skill to apply it in the experimental situation. There is probably no testing of causal laws which does not rely on causal knowledge (both scientific and prescientific). And there is hardly any individual item of scientific causal knowledge which is not “embedded” in a *system* of such knowledge.

VI

Our knowledge of a causal relation may have been obtained by successfully testing a *hypothesis*. It may also happen that a causal relation which was thought to be known later becomes subject to *doubt*.

A primitive or prescientific idea about a causal connection may become corrected with the advancement of (scientific) knowledge. There is (or was) a popular belief that being exposed to a cold temperature might cause a “cold”. We now know that the symptoms of the illness are caused by bacteria—and that the cooling of the body is only a circumstantial condition under which the working of the bacteria on the body becomes “efficacious”.

Getting cool/catching a cold is a very “rough” “uniformity of nature”. Shall knowledge of it count as “causal knowledge” at all? What is here “common knowledge” is something like the following: under certain circumstances, letting oneself become cool easily results in a cold. If one had to characterize the circumstances in greater detail one could say, for example: when there is already a cold “about the place”.

Suppose the circumstances are such that if a certain man exposes himself to a cold temperature, he will get a cold. (The germs are already in his body “awaiting an opportunity” to attack him.) He then gets cool—and a cold. Shall we say that the cooling, i.e. the drop in temperature of his body, was the cause of the cold? I think it is perfectly correct to say this.

But is this not to turn things upside down? Was not the

"real" cause internal processes in the person's body, processes which in their turn were initiated by the germs? The drop in temperature was only a "condition" which, when it was satisfied, made the real cause "operative", i.e. productive of the malaise.

The distinction between "cause" and "condition" is familiar. And it is, for many purposes, a useful distinction to make. But what counts as condition and what as cause is not fixed "in the nature of things". The distinction is relative. Given the conditions under which a temperature-drop will make operative germs already present in the body, the temperature-drop is a cause and the presence of germs a condition. If, however, the germs are not yet in the body, but the conditions under which they would cause a cold are satisfied, then it is the infusion of germs which is the cause.

There may be reasons for calling the germs a "more real" cause of a cold than a drop in body-temperature: for example, that the germs can also be activated when no drop in temperature takes place, whereas a change in body-temperature cannot produce a cold unless there are germs. But this would not run contrary to the fact that there are circumstances under which it is perfectly correct (and not "unscientific") to say that a cold was caused by a drop in temperature of the body.

The reason for calling the temperature-drop "cause" should also be plain: it is the *change* which, under the circumstances, we hold responsible for another *change* (the "outbreak" of the cold). If, on another occasion, the cooling of the body does not result in a cold, we should say that the circumstances were not the "right" ones. For, we *know* that a temperature-drop *can* cause a cold, i.e. *will do it* under appropriate circumstances. This is "common knowledge".

VII

In an important type of case, to think that the happening of *c* is a cause of the happening of *e* commits one to holding that, "*ceteris paribus*", the happening of *c* will (always) be accompanied by the happening of *e*. Moreover, this "will be" is more than a statement about what will happen (*e*), *if something else* happens (*c*) under appropriate circumstances (*C*). In a charac-

teristic sense, the "will be" also covers all "past futures" and all "future pasts". This means the following:

Of all past occasions when the circumstances *C* prevailed but *c* did not happen and of all future occasions when the circumstances *C* will have prevailed but *c* will not have happened, it is true that, *had c* been there on those occasions, *c* *would have* accompanied it.

It seems to me that it is in this implicit commitment to a counterfactual conditional assertion that our belief in the causal efficacy of one event upon another, or in the "causal bond" linking two events, consists.

VIII

If it is true that the differentiating mark between a causal bond and an accidental concomitance is that the former but not the latter supports counterfactual conditionals, then the question how one acquires causal knowledge is essentially the question how one can get to know, if at all, the truth of counterfactual conditionals.

Counterfactual conditionals are, one could say, *retrospective* statements. They speak about what would have happened, had something been different from what it actually was, is, or will have been. In the case of causal counterfactuals, the actuality is that, on some occasion or succession of occasions, the cause-event does not happen. The contrasting non-actuality is that both the cause- and the effect-event happen. In order to "verify" the counterfactual statement we ought somehow to make the actual and the non-actual "change place". How can this be done?

Literally this can of course not be done at all. Saying that the actual (factual) and non-actual (counter-factual) change place is a metaphor. But it may be a useful *façon de parler* in speaking about things which literally can and do take place.

As far as I can see, the acquisition of causal knowledge presupposes that there are situations in which we are *certain* that the cause- and the effect-events *will not* occur although they *can* occur. If the cause-event is something we can make happen, then, by producing it, we can actualize what otherwise would have remained unactualized. Assume now that in such a situation we

make the cause-event happen. If, having done this, we find that the effect-event does not happen, we may conclude either that the presumed cause is not really a cause with that effect, or that the *circumstances under which it is efficacious* are not satisfied—or we may suspend judgment. If, however, we observe the effect-event we also know that, had the cause-event *not* occurred on that occasion, then it would have been true retrospectively to maintain that, *if it had occurred*, the effect-event would have accompanied it. We, as it were, “proved” this by making the cause-event occur and observing what happened then. Metaphorically speaking, we proved it by making the actual and the non-actual “change place”.

IX

But could we not simply *wait* (“passively”) for the cause-event to occur and then, when it occurs and is followed by the effect-event, gain the same insight into the counter-factual truth as we get from the “experiment”? I think the answer is “No.” Mere observation of regular sequences in nature may suggest to us the existence of causal connections and may make us put forward various causal conjectures or hypotheses. Further observations may confirm or refute such hypotheses. Perhaps after long confirmation we say we “know” such a hypothesis to be true. It would be futile, I think, to dispute whether this can be *genuine knowledge*, or not. But it is important to see that and why the possibility of an experimentalist interference with the case changes the epistemic situation radically—not only in degree but in conceptual character. The following considerations should help us see this more clearly:

Suppose we are familiar from experience with a regular sequence: c_1 followed by c_2 followed by e . What sort of causal connectedness might this suggest? There are two possibilities. One is that c_1 causes c_2 which in turn causes e . The other is that c_1 is the cause of the sequence: c_2 followed by e . The problem connected with coming to know the second possibility is the same as the problem of coming to know a simple causal relation of the type: c causes e . The problem, again, of coming to know that there is a causal chain, c_1 causes c_2 which causes e is more

complicated. It can be split up in two. The first is to come to know that c_1 causes the sequence: c_2 followed by e . On this we already commented. The second is to come to know that c_2 , by itself, causes e . To this end we must study cases in which c_2 occurs, but *not* as an effect of c_1 , nor of any other known cause of its occurrence. Because if the sequence c_2 followed by e is an effect of, say, c_1' then we are again faced with the problem of "detaching" c_2 from this cause in order to find out whether c_2 , *by itself*, is causally efficacious. Thus in order to test the causal efficacy of c_2 there must exist, or we must by manipulation be able to secure the existence of, situations such that the introduction of c_2 into them is in our control, i.e. such that we are confident that c_2 and e will not make their appearance in them unless made to appear. If we can make and do make c_2 happen in such a situation and find that it is not followed by e , we may conclude that c_2 has not the causal power of producing e , at least not under the circumstances accompanying the experiment. If, again, e follows upon the appearance of c_2 , we have confirmed that there is a causal connection between the two factors—as distinct from a mere concomitance due to the existence of a common cause for both of them. Because we are now entitled to say that had we let c_2 remain absent on the occasion when we made it happen, then it would have been true that *had* we made it happen, e *would have* followed. We "proved" this by intervening with that which we were certain would otherwise have taken place.

Our certainty *may*, of course, have been "deceptive" in the sense that the result of our intervention was, in fact, due to some cause external to us. But in order to find this out and come to regard this external factor as a cause of the sequence c_2 followed by e , we should again have to go through the same epistemic procedure for coming to hold another counterfactual conditional true.

Two very simple examples will illustrate these abstract lines of thought:

I put fire to a sheet of paper. The edge of the paper turns first brown then black, the sheet crumbles, and finally turns to ashes. The stages in this process are, broadly speaking, successive

effects of *one* cause, viz. the heat (the flame) to which the paper is being exposed. It is not, for example, the change of color at the edges which makes the paper crumble. How do I know this? Nothing is easier: I can give to a sheet of paper those successive colorings without producing the subsequent effects of the heat when it devours the sheet. And I can crumble the paper without turning it to ashes.

A stone hits a window and breaks it. Air enters the room from outside and there is a drop in indoor temperature. In this chain of events every link is a cause of the next one. It is not the hit of the stone against the glass which *first* breaks the window and *eventually* cools the room. It is the entering of outside air into the chamber which has the cooling effect. This is easily established—"experimentally" if needed.

X

Our "common knowledge" of causes—such as that water boils when heated—is founded in man's accumulated experience about what follows (the effect) when a certain thing happens (the cause) under *familiar* circumstances. Such knowledge is often, perhaps usually, intimately interwoven with our practical life, i.e. with our ability to effect changes by doing other things under those circumstances—for example, to make water boil by heating it. When such a connection with manipulation is missing—as in the case of lightning and thunder, say—a conjectured causal connection is, at the prescientific stage, often associated with ideas of a being endowed with superhuman powers, for example a thunder-god. At a scientific stage, such conjectures are associated with a systematic search for experiments, i.e. for *learning to reproduce the cause-event under controlled circumstances*.

Why is it "irresistible" to think of lightning as the cause of thunder although the connection is not very similar to connections familiar from *our* practical life? This is worth reflecting about. I suppose that one reason is that lightning is a striking intervention with an existing state in nature. It is an event the (natural) cause of which is, at the prescientific stage, completely

hidden from our knowledge. In this the occurrence of lightning resembles things the production of which is "in the hands" of an agent.

XI

A cause, then, is something the occurrence of which *initiates* a sequence of events (also) when it is not itself the effect of another cause. The occurrence of a cause-event may of course be embedded in a *causal chain*, i.e. occur as the effect of another cause. But any later link in such a chain can be known to be (itself) causally efficacious only by being detached from the preceding links and made to occur as "initiator" of the succeeding part of the causal sequence.

If I am right in thinking that causal concatenations can be known only as detachable parts of bigger wholes within which non-caused initiation of changes is taken for granted, then the truth of *determinism* can at most become established for fragments of the world and not for the world as a totality. But may not determinism nevertheless be *true* for the totality—only we cannot come to *know* its truth? The question seems to me idle. How would this truth "manifest" itself? It manifests itself to the extent that we get to know causal connections between events of given generic character. *Belief* in determinism may influence our orientation in the world and direct our research. It can function as a constant urge to search for causes. But that determinism is true cannot itself be "causal knowledge".

Scepticism and Sanity¹

HIDE ISHIGURO

I. The Problem

The discussion of the question "How can I be sure that I am not now mad?" has recently become controversial in a new way, as seeming to acknowledge that there is a thing called madness. People with unrelated abnormalities have in the past been labelled mad, and because there is social discrimination against people who were thus labelled, some have claimed that insanity is nothing but a social or political category. The way we treat those we consider mentally ill naturally reflects our dogmas and social conventions. Influential thinkers such as Laing in England, Deleuze in France, or Szasz in the U.S.A. have concluded from this that there is no fundamental distinction between those labelled mentally ill, and the rest of us (apart from the fact that they are thus labelled). This has meant that, although philosophers have become more sensitive to the unwarranted assumptions and values reflected in the way the institutions which cater for the so-called "mad" are run (and surely this new sensitivity is a change for the good), they have become circumspect about using sanity as a normative concept. Even to mention the mentally ill in epistemological contexts lays one

1. Some parts of this article developed out of a paper read to the A.P.A. in December 1975, where I reached a different conclusion, and on which occasion I profited from comments made by Annette Baier. I have since profited from discussion with Doris Baker and Myles Burnyeat, and from suggestions by the editors of this volume.

open to accusations of using discriminatory categories, of defending the *status quo*.

However unfortunate this may be, we do encounter people who manifestly think and yet think along trains of thought we do not understand. It is not so much that we do not see why such people think in the way they do and that we disagree with them; rather we are at a loss to know what their thinking is; we cannot put ourselves in their shoes. (Among those regarded as mentally ill, it is only the people whose cognitive powers seem to be impaired in such a way that we fail to see how their beliefs relate to reality who concern us here.) Can this fact be sensibly questioned? Can we coherently suppose that we, the seemingly ordinary people, are really the ones who are insane? My answer to this question is: no, unfashionable as it may appear. This paper is an attempt to show why we cannot replicate for the case of our sanity the philosophical doubts that we can formulate for our senses. This thought-experiment does not concern cases where my experiences are simply odd (e.g. when I hear voices when no one seems to be around, or people appear to me to be so different from what I have taken them to be in the past that I wonder whether I am mad). Such a case is like wondering whether something is wrong with my eyes, and wondering whether I am having double vision, when objects seem to duplicate themselves, or wondering whether I am astigmatic when the shape which I recall was regular suddenly appears to me distorted. My claim concerns rather the possibility of doubting my sanity even when my experience seems no more abnormal than at other times. I will try to show that to conceive of myself as possibly being mad in the pertinent manner now is to imagine myself to be in a condition where I should abandon the trust I have in my ability to tell what I am perceiving in standard situations like this. I can give content to this only if I conceive of myself as being now in a condition such that in principle I could come to realize that I had been mad. In carrying out a thought-experiment of this kind, I hope not to infringe two precepts whose importance Norman Malcolm has brought out so admirably in his writings: not to be taken in by false analogies, and not unwittingly to destroy the framework in which alone the very issue can be meaningfully raised.

The table and chair before me appear to me to be of the same colour. I may wonder whether I am colour-blind and they are really of two quite different colours. Again, I may wonder whether I am making a misjudgment not through physical handicap but through carelessness. Is what I see as elliptical in the distance actually *circular*? If it makes sense to wonder this, it makes sense to suppose this. Many have thought that I can also wonder whether I might be dreaming. Would it be possible for me to ponder and then judge that I may eventually wake up and realize that I had been asleep, and not, as I supposed, thinking about scepticism sitting before a table? This has been denied by Professor Malcolm in his celebrated book on *dreaming*. Whether we take his position or not, we can at least agree that I can sometimes wonder whether I *have* been dreaming. Did I really see those people and hear those odd words, or did I doze off and dream those things? Am I mistaken about my experiences of the immediate past? It seems quite clear what it is to wonder about this.

I can, in a manner similar to this, wonder whether I had been mad. "You mean," Gilbert Pinfold says, "*that everything I've heard said, I've been saying to myself? It's hardly conceivable.*" Can I, however, wonder whether here and now I am mad and am a victim of an insanity affecting my cognitive powers, having all the experiences and beliefs I have? Would it be to suppose that I am the victim of hallucinations? Can one say of a thought or a belief on its own that it is mad or sane in the same way that one can say of it that it is true or false? Or is "mad" a feature like "consistent" or "stray"—a feature of a thought in relation to other thoughts? Or is it a feature of a set of beliefs or thoughts like "orderly" or "coherent"?

II. Sanity and Cartesian Doubt

In the First Meditation where Descartes expounds his famous methodological doubt, he raises all the three doubts just mentioned, the doubt about his senses, the doubt whether he is dreaming, and the doubt about madness. Since then, many people have discussed his scepticism about the senses and the doubt whether he may be dreaming. Apart from Wittgenstein's

raising a very similar problem in his *On Certainty*, very few have discussed Descartes's hypothesis about the possibility of one's being mad.² As I hope to show later, the problems involved are related, however, to the question of "Radical Interpretation," discussed extensively in recent years by Donald Davidson and others.

Let me begin by reminding you how Descartes introduces the hypothesis of madness in the course of his systematic doubt. Descartes suggests that, although the senses may at times deceive us about some minute or remote objects, yet there are many facts we learn from the senses about which doubt is plainly impossible. As Moore and Wittgenstein were also to do later, Descartes takes the existence of his hands in front of him and his body as the paradigm of things he cannot doubt. It is at this point that he briefly considers the hypothesis of his being mad.

Unless indeed I likened (*comparare*) myself to some lunatics (*insani*) whose brains are so upset by persistent melancholy vapours that they firmly assert they are kings, when really they are miserably poor, or that they are clad in purple when they are really naked; or that they have a head of pottery, or are pumpkins, or are made of glass; but then they are madmen (*amentes*) and I should appear (*viderer*) no less mad (*demens*) if I were to try to transfer their example on to myself (*si quod ab iis exemplum ad me transferrem*).

Thus Descartes first introduces being mad as a state of certain people he observes in the world. He has a causal theory about the physiology of the madmen, but for him the criterion

2. The only recent examples I know are those of Michel Foucault in three pages of his *Madness and Civilization* (*Histoire de la Folie*, Paris: Plon, 1961; 2nd ed., Paris: Gallimard, 1972, pp. 56-59, and also App II, pp. 583-603); Jacques Derrida in an article commenting on Foucault called "Cogito et Histoire de la Folie" (in *Ecriture et la Différence*, Paris: Editions de Seuil, 1967); a short discussion in Professor Harry Frankfurt's book, *Demons, Dreamers and Madmen* (New York: Bobbs-Merrill, 1970); and in an unpublished paper by Professor André Gombay which recently came to my notice. Brief, passing comments on the issue may be found in Dr. Anthony Kenny's book, *Descartes* (New York: Random House, 1968). My attention has just been drawn to Steven De Haven's article in *Analysis*, March 1978, but he does not seem to think that there is any special difficulty about the madness hypothesis either for Descartes or for himself.

of madness lies in what they say and do. *Were* I to suppose that I have no body or to suppose that my hands don't exist, I should be behaving like them. But Descartes puts aside the hypothesis, because to suppose that I could do what madmen appear to be doing is to appear just as mad as they. And, instead of pursuing the case further, he turns immediately to the supposition that one may be dreaming, saying that in sleep he has "the same kind of impressions as those which lunatics have when awake, or even ones that resemble reality even less".

Now, why does Descartes cut short the consideration of the supposition that he might be mad and then proceed to wonder whether he is dreaming? Some have said that the reason lies in the fact that for Descartes the two suppositions are basically the same. Dr. Anthony Kenny takes this position³ and goes on to say that, although the question "How do I know that I am not mad?" is one of great philosophical interest, it is not pursued by Descartes, possibly because it might seem offensive to the reader, but is replaced by the sceptical doubt about dreaming.⁴ Derrida too suggests that the hypothesis about madness and about dreaming are of fundamentally the same nature, dreaming being only more extreme in its delusory character. He does not think that replacing one by the other makes any important difference for Descartes.

Some have seen a more philosophical point in Descartes's cursory rejection of the madness hypothesis. It has been claimed that Descartes has no reasonable alternative. As the whole point

3. Anthony Kenny, *Descartes*, p. 15.

4. Indeed in another work of Descartes, *The Search after Truth*, one of the protagonists says, "Since it is not sufficient for me to tell you that the senses deceive us in certain cases where you perceive, in order to make you fear being deceived by them on other occasions when you are not aware of it, I shall go further and ask if you have ever been a melancholic man of the nature of those who believe themselves to be vases, or who think some part of their body is of enormous size; they would swear that they see and touch that which they imagine they do. And it is true that any ordinary man would be indignant if anyone were to say to him that he could not have any more reason than they to be certain of his opinion, since it rests equally with theirs on what the senses and his imagination represent to him. But you cannot be annoyed if I ask you whether you are not like other men subject to sleep." (Descartes obviously thinks that, philosophically speaking, the ordinary man has no right to be indignant.)

raising a very similar problem in his *On Certainty*, very few have discussed Descartes's hypothesis about the possibility of one's being mad.² As I hope to show later, the problems involved are related, however, to the question of "Radical Interpretation," discussed extensively in recent years by Donald Davidson and others.

Let me begin by reminding you how Descartes introduces the hypothesis of madness in the course of his systematic doubt. Descartes suggests that, although the senses may at times deceive us about some minute or remote objects, yet there are many facts we learn from the senses about which doubt is plainly impossible. As Moore and Wittgenstein were also to do later, Descartes takes the existence of his hands in front of him and his body as the paradigm of things he cannot doubt. It is at this point that he briefly considers the hypothesis of his being mad.

Unless indeed I likened (*comparare*) myself to some lunatics (*insani*) whose brains are so upset by persistent melancholy vapours that they firmly assert they are kings, when really they are miserably poor, or that they are clad in purple when they are really naked; or that they have a head of pottery, or are pumpkins, or are made of glass; but then they are madmen (*amentes*) and I should appear (*viderer*) no less mad (*demens*) if I were to try to transfer their example on to myself (*si quod ab iis exemplum ad me transferrem*).

Thus Descartes first introduces being mad as a state of certain people he observes in the world. He has a causal theory about the physiology of the madmen, but for him the criterion

2. The only recent examples I know are those of Michel Foucault in three pages of his *Madness and Civilization* (*Histoire de la Folie*, Paris: Plon, 1961; 2nd ed., Paris: Gallimard, 1972, pp. 56-59, and also App. II, pp. 583-603); Jacques Derrida in an article commenting on Foucault called "Cogito et Histoire de la Folie" (in *Ecriture et la Différence*, Paris: Editions de Seuil, 1967); a short discussion in Professor Harry Frankfurt's book, *Demons, Dreamers and Madmen* (New York: Bobbs-Merrill, 1970); and in an unpublished paper by Professor André Gombay which recently came to my notice. Brief, passing comments on the issue may be found in Dr. Anthony Kenny's book, *Descartes* (New York: Random House, 1968) My attention has just been drawn to Steven De Haven's article in *Analysis*, March 1978, but he does not seem to think that there is any special difficulty about the madness hypothesis either for Descartes or for himself.

of madness lies in what they say and do. Were I to suppose that I have no body or to suppose that my hands don't exist, I should be behaving like them. But Descartes puts aside the hypothesis, because to suppose that I could do what madmen appear to be doing is to appear just as mad as they. And, instead of pursuing the case further, he turns immediately to the supposition that one may be dreaming, saying that in sleep he has "the same kind of impressions as those which lunatics have when awake, or even ones that resemble reality even less".

Now, why does Descartes cut short the consideration of the supposition that he might be mad and then proceed to wonder whether he is dreaming? Some have said that the reason lies in the fact that for Descartes the two suppositions are basically the same. Dr. Anthony Kenny takes this position³ and goes on to say that, although the question "How do I know that I am not mad?" is one of great philosophical interest, it is not pursued by Descartes, possibly because it might seem offensive to the reader, but is replaced by the sceptical doubt about dreaming.⁴ Derrida too suggests that the hypothesis about madness and about dreaming are of fundamentally the same nature, dreaming being only more extreme in its delusory character. He does not think that replacing one by the other makes any important difference for Descartes.

Some have seen a more philosophical point in Descartes's cursory rejection of the madness hypothesis. It has been claimed that Descartes has no reasonable alternative. As the whole point

3. Anthony Kenny, *Descartes*, p. 15.

4. Indeed in another work of Descartes, *The Search after Truth*, one of the protagonists says, "Since it is not sufficient for me to tell you that the senses deceive us in certain cases where you perceive, in order to make you fear being deceived by them on other occasions when you are not aware of it, I shall go further and ask if you have ever been a melancholic man of the nature of those who believe themselves to be vases, or who think some part of their body is of enormous size; they would swear that they see and touch that which they imagine they do. And it is true that any ordinary man would be indignant if anyone were to say to him that he could not have any more reason than they to be certain of his opinion, since it rests equally with theirs on what the senses and his imagination represent to him. But you cannot be annoyed if I ask you whether you are not like other men subject to sleep." (Descartes obviously thinks that, philosophically speaking, the ordinary man has no right to be indignant.)

of Descartes's critical examination of his former opinions lies in determining whether there are reasonable grounds for doubt, Descartes cannot carry out his task without assuming his own sanity—even if the assumption be provisional and heuristic.⁵ It has also been suggested that Descartes cannot envisage such a possibility in the way he seriously envisages the possibility of the malicious demon who systematically deceives us. For the malicious demon universally affects all human minds, whereas madness is something which affects only some men. But is *Meditations* devoted as Professor Frankfurt has said to the question: "How, if at all, can certainty be attained by those who are best qualified to pursue it?"

To take the second point first, I cannot see why it is proper for Descartes (or for anyone engaged in a systematic doubt) to suppose himself to be free from all personal deficiencies—since after all he takes seriously the possibility of his senses deceiving him, and there may be deficiencies of the senses which affect some men and not all others. If *Meditations* is committed to doubting everything that it is reasonable to doubt, then it must be reasonable to wonder whether one has certain personal deficiencies. When one carries out a visual perception, one cannot assume that one is not color-blind. One has to have a method which will enable one to discover that one is color-blind, if one is. If its purpose were to rule out personal defects then Descartes's rejection of the madness hypothesis would be quite unjustified. Frankfurt's first point is more relevant, but it does not seem to be formulated in quite the right way. In what sense does Descartes have no alternative but to replace one doubt by the other? Frankfurt writes as if I could freely assume either that I am mad or that I am sane but, because it is the only one that enables me to carry out what I want to do, I choose the latter assumption. What I want to argue is that there is even a difficulty about giving substance to the supposition that one is now mad, and that the only way I can do this is, as in the case of dreaming, to think that in principle there could be a future state in which I was certain that I had been mad and rightly so.

5. Harry Frankfurt, *Demons, Dreamers and Madmen*, p. 38.

III. Irrational Thinking

Before entering the argument, I would like to mention a mistake which recurs in recent discussions about madness and irrationality. This is the view that mad thoughts can, and rational thoughts cannot, be explained causally in ways making no reference to the truth of the other thoughts or beliefs a person has. Rational thoughts, it has been said, presuppose freedom—freedom to change one's mind given certain evidence—and are linked with other beliefs and other thoughts of the person, whereas mad thoughts, it is suggested, are not. In *On Certainty*, sections 73-74, Wittgenstein wonders what makes us treat a belief either as a mistake, or as a mental disturbance, and asks whether the difference lies in a mistake having a reason (*Grund*) as well as a cause. But the beliefs and actions of the mad are quite often related to, and supported by, what they take to be (and sometimes are) other true beliefs. Frequently they have their reasons for holding on to their beliefs. Conversely, if there are neurophysiological patterns corresponding to thought processes of madmen, then there are patterns of neurophysiology of normal thoughts as well. If laws of rational thinking cannot be reduced to causal laws governing phenomena of neurophysiological categories, and if this shows that a theoretical reductionism is impossible, then neither can processes of thought that are irrational be reduced to specific causal laws governing neurophysiological phenomena. There is no theoretical reduction here either. This is the case even if certain physiological conditions are sufficient for causally bringing about a mental derangement.

Another point which should be made is that recent discussions about rationality have too often been concerned with rationality as a capacity which distinguishes humans from other animals. This has the effect either of leaving no scope for men who think but lack rationality, or of making both rationality and irrationality exclusively ascribable to beliefs of sane men—arbitrarily omitting the insane from discussion about beliefs and desires. For example, Professor Jonathan Bennett once wrote that he used "rationality" to mean "whatever it is that humans possess which marks them off, in respect to intellectual capacity,

sharply and importantly from all other known species." (*Rationality*, p. 5) It seems misleading to call this capacity "rationality." For Descartes, what distinguished men from all others was the fact that they thought. But even for him, having thoughts (*cogitationes*) or being aware was not the same as being rational. Rationality, or common sense ("*le bon sens*") was, as he says at the beginning of the *Discourse*, one of the properties that are distributed most widely among men—but he does not say that one *cannot* be an irrational thinking being. Madmen obviously have thoughts, and that meant for Descartes that they had a soul. A madman's status is quite different from the other animals, which Descartes considered to be complex machines. And even if we reject Cartesian dualism, we can still agree that insanity is a condition which only men are able to fall into, just as the possibility of rationality is their normal condition. We can understand people as having beliefs and thoughts without being able to understand how they arrive at or can hold on to such thoughts or beliefs. To suppose that one is now mad is not like supposing oneself to be a table, or even a bat. If I am not mad this is a mere contingent fact. My being mad is a physical possibility. Why then is there any difficulty in supposing that I am mad now? (The fact that "mad" has been a colloquial term which was applied to disparate groups of people need not concern us here. In trying to suppose myself to be mad, I am trying like Descartes, to suppose myself to be a member of that subgroup of people whose cognitive link with reality seems to have broken down.)

IV. Truths and Beliefs

It is very important to continue to remember here that what is in question is not whether I could imagine myself having the experiences of the mad, of hearing voices, or feeling persecuted, but whether I could conceive that now, when my experiences appear to be normal, I may be mad. Here are my hands in front of me. To doubt this would be to admit that it makes sense to conceive of there being facts that would render uncertain my belief that I do have hands. But as Wittgenstein points out in *On Certainty*, if I am not to trust my judgment about my hands

in front of me why would I trust my judgment about the other fact? Yet what Descartes was asking us to do was to try to conceive that our actual present experience was that of dreaming or of being mad.

It is not enough for me to suppose that I have all the experiences I now have, and that other people *call* me mad. They may be falsely ascribing madness to me, whereas my task was to suppose myself to be truly insane while having all the experiences I do have. Neither is it enough to suppose that some of the things that seem to me to be the case are not actually the case. This would just be to suppose myself having some false beliefs, and as Descartes himself acknowledged, there is nothing impossible or mad about just that. A figure in the back of the garden which I take to be a human being, may be a life-size wax doll, installed to fool me. The person passing by who appears menacing may in fact be gentle. I can have false beliefs for good reasons, but also for bad reasons, without being mad.

The difference between the madmen described by Descartes who believe themselves to be made of porcelain, and myself, is not *just* a disagreement in certain opinions. For we do not consider mad all people with whom we disagree, even when the beliefs are quite fundamental ones that affect the way which we see many things around us. Indeed, a madman may have many true beliefs, and it often seems that he perceives many things quite correctly. Certainly we cannot assume that he perceives *everything* wrongly. Thus, to suppose oneself to be mad is not to imagine that every belief one has is false or that every perception one has is distorted.

Most thoughts or desires are not mad or rational in and of themselves. The very same thoughts could be reasonable thoughts to hold in the presence of certain kinds of evidence and certain sets of beliefs (which, even if false, are not necessarily mad in themselves); and they could be mad thoughts to hold in the presence of a different kind of evidence and different sustaining beliefs. Primarily then it is to a person or a person's way of maintaining a system of beliefs and desires over a certain period of time that we ascribe madness. It is not to a belief or a desire on its own.

Now the beliefs of people whom we consider sane are not

sharply and importantly from all other known species." (*Rationality*, p. 5) It seems misleading to call this capacity "rationality." For Descartes, what distinguished men from all others was the fact that they thought. But even for him, having thoughts (*cogitationes*) or being aware was not the same as being rational. Rationality, or common sense ("*le bon sens*") was, as he says at the beginning of the *Discourse*, one of the properties that are distributed most widely among men—but he does not say that one *cannot* be an irrational thinking being. Madmen obviously have thoughts, and that meant for Descartes that they had a soul. A madman's status is quite different from the other animals, which Descartes considered to be complex machines. And even if we reject Cartesian dualism, we can still agree that insanity is a condition which only men are able to fall into, just as the possibility of rationality is their normal condition. We can understand people as having beliefs and thoughts without being able to understand how they arrive at or can hold on to such thoughts or beliefs. To suppose that one is now mad is not like supposing oneself to be a table, or even a bat. If I am not mad this is a mere contingent fact. My being mad is a physical possibility. Why then is there any difficulty in supposing that I am mad now? (The fact that "mad" has been a colloquial term which was applied to disparate groups of people need not concern us here. In trying to suppose myself to be mad, I am trying like Descartes, to suppose myself to be a member of that subgroup of people whose cognitive link with reality seems to have broken down.)

IV. Truths and Beliefs

It is very important to continue to remember here that what is in question is not whether I could imagine myself having the experiences of the mad, of hearing voices, or feeling persecuted, but whether I could conceive that now, when my experiences appear to be normal, I may be mad. Here are my hands in front of me. To doubt this would be to admit that it makes sense to conceive of there being facts that would render uncertain my belief that I do have hands. But as Wittgenstein points out in *On Certainty*, if I am not to trust my judgment about my hands

in front of me why would I trust my judgment about the other fact? Yet what Descartes was asking us to do was to try to conceive that our actual present experience was that of dreaming or of being mad.

It is not enough for me to suppose that I have all the experiences I now have, and that other people *call* me mad. They may be falsely ascribing madness to me, whereas my task was to suppose myself to be truly insane while having all the experiences I do have. Neither is it enough to suppose that some of the things that seem to me to be the case are not actually the case. This would just be to suppose myself having some false beliefs, and as Descartes himself acknowledged, there is nothing impossible or mad about just that. A figure in the back of the garden which I take to be a human being, may be a life-size wax doll, installed to fool me. The person passing by who appears menacing may in fact be gentle. I can have false beliefs for good reasons, but also for bad reasons, without being mad.

The difference between the madmen described by Descartes who believe themselves to be made of porcelain, and myself, is not *just* a disagreement in certain opinions. For we do not consider mad all people with whom we disagree, even when the beliefs are quite fundamental ones that affect the way which we see many things around us. Indeed, a madman may have many true beliefs, and it often seems that he perceives many things quite correctly. Certainly we cannot assume that he perceives *everything* wrongly. Thus, to suppose oneself to be mad is not to imagine that every belief one has is false or that every perception one has is distorted.

Most thoughts or desires are not mad or rational in and of themselves. The very same thoughts could be reasonable thoughts to hold in the presence of certain kinds of evidence and certain sets of beliefs (which, even if false, are not necessarily mad in themselves); and they could be mad thoughts to hold in the presence of a different kind of evidence and different sustaining beliefs. Primarily then it is to a person or a person's way of maintaining a system of beliefs and desires over a certain period of time that we ascribe madness. It is not to a belief or a desire on its own.

Now the beliefs of people whom we consider sane are not

merely generated or supported by their grounds of truth. People's desires and emotions, expectations and imaginations, affect the beliefs they hold. That this is so does not make people incomprehensible. Indeed we can notice such traits in ourselves. And we can even make sense of the systems of belief, and attitudes of people for whom we have no sympathy, or with whom we disagree strongly over many issues, or even who appear alien to us, by seeing what their desires and expectations are. Nevertheless so long as we are sane, the concern for the truth of our beliefs is in some way always there for us. We have somehow to put up with and manage the known irrationality of some of our beliefs and attitudes.

In people whom we do not consider mad we often see that there are isolated systems of beliefs which are somehow detached from reality or are not effectively linked with their actions, as in the cases of weakness of will or self-deception. In such cases, however, we detect a tension and an unstable state in the agent. The tension corresponds to his ability to be concerned about the truth of his beliefs. The world impinges on a person's beliefs and even where the way in which we measure or establish facts depends on conventions these are not arbitrary. That is why in interpreting other people we in general adopt what Richard Grandy has called the *Principle of Humanity*: the principle that the imputed pattern of relations among beliefs, desires, and the world be as similar to our own as possible.⁶

For example, in order to guess the meaning of an utterance of a foreign speaker, no one would merely observe the speaker's environment in order to conjecture what his beliefs might be. We might see that his physical environment is uncomfortable and bleak, but also know from his past and his other behavior that he desires hardship and obtains satisfaction from discomfort, and that this is linked with his views about afterlife, etc. What we then do is to think what we would believe and say, if we had the same desires as he and found ourselves in the same environment as his. At the same time, his behavior, including his verbal behavior, is part of what makes us ascribe certain short-term reasons as well as desires to him. (From what I have

6. Richard Grandy, "Reference, Meaning and Belief," *Journal of Philosophy*, 70, no. 14 (Aug. 16, 1973).

said previously, it should be clear that I do not treat verbal behavior in quite the same way as other behavior.) If I were to behave as he does what reasons would I have for doing so? If, e.g., I were to behave in such an unperturbed and cheerful manner in such an uncomfortable situation, would I not have a reason to conceal my discomfort? And what would my reason likely be?

To judge a man mad is to avow the breakdown of the *Principle of Humanity*. It is perhaps when the kind of tensions that are created to cope with our irrationalities seems to be absent that we acknowledge the existence of people whose systems of belief we fail to penetrate. We do not see how they relate or fail to relate to the world outside them. Even if we can recognize from the outside certain patterns in such a person's train of thought, what we cannot understand is why the thinker takes these patterns to correspond to a line of thinking which leads him to true beliefs. (To say this is not like saying of a person that his logic is intuitionistic rather than classic. For even a classical logician can understand the rules of intuitionistic logic, and see what makes an intuitionist believe that mathematical thinking should be governed by such rules.)

Many theorists have tried to show that it is more possible than is generally supposed to understand people who are diagnosed as psychotic. But it is important to distinguish between the understanding in the sense of seeing the desires or emotions that led people to construct certain systems of beliefs, and the understanding in the sense relevant to our question how they manage to hold on to the beliefs they have, despite all the counterevidence that to us outside observers seems to be available to them.

V. Insanity and Language

When we fail to understand people's beliefs or people's acts, it is tempting to say that their language is different from ours. We see the madman hold his hands in front of him and say "I have no hands". But the fact that we observers can see that he does have hands and the fact that it is obvious both that he is looking at his hands and is not lying will not lead us to

conclude that he understands the predicate "has hands" to mean something different from what we take it to mean. It will not do to say that his concepts are different from ours or that he uses words differently. We cannot apply principles like the Principle of Charity piecemeal, even the so-called "improved" Principle of Charity, i.e. the principle to interpret in such a way as to maximize the agreement of the beliefs of the person whose verbal behavior I am interpreting and the beliefs I would have had, if I had behaved as he did. No such principle licenses the simple inference from "I don't understand him" to "He speaks a different language". People have too often been tempted to do this in writing about the language of schizophrenics. People assume too easily that words have lost their customary meaning for them. But when the schizoid says that thoughts which are not his own get into his mind, the experience he is describing is abnormal precisely because he means by "not his own" what is customarily meant by these words.

Donald Davidson has rightly stressed that a theory of meaning cannot be given independently of a theory of belief.⁷ Descartes's madman is probably not in doubt that a condition obtains which might appear to be that of his having hands. His language is that of our community. When he says of other people that they have hands, he presumably means what we mean by having hands. He may have a whole story why he himself really has no hands. For example, what appear to be his hands are false hands. He has lost his real hands since he committed the disgusting act, and so on. The fact that a person applies words wrongly to things, or interprets phenomena as symptoms of a certain non-existent state of affairs *p* does not show that he uses words with different meanings, or that he has a different concept of *p*. Even the bizarre rambling sentences uttered by a schizophrenic are taken as expressing an abnormal experience precisely because the words with their customary meanings are used in unexpected ways.

Neither can we assume that when we fail to understand a particular belief of a madman it is because the belief is self-contradictory or incoherent. Descartes's madman looks at his

7. Donald Davidson, "Radical Interpretation," *Dialectica* (1973).

hands and says "I have no hands". His belief is not that "my hands which I am looking at are not my hands". When I describe what he is doing and say "He looks at his hands, and yet believes that he has no hands", the occurrence of the words "his hands" is referentially transparent. The thought he has that his hands don't exist, and his act of denying that these are his hands, are not expressions of a belief that his hands are not his hands.

As many psychiatrists have reported, some people who are diagnosed as schizophrenics invent new words and use them in ways which are difficult to follow. In such cases it is obvious that their language is not quite the same as ours. When we fail to understand sentences which they use containing these new words, our failure to understand the meaning they attach to the words they have invented, and our failure to identify the content of their beliefs, come together. We find it difficult to translate what they are saying into our own language because it is so hard to establish what exactly it is that they believe to be true.

Yet this is not invariably the case with people we consider to be mad. A person's way of holding on to, or arriving at, false beliefs (rather than a contradictory belief, or one whose expressions are verbally incomprehensible to us) is often enough to make us unable to understand a person's system of thinking. And this means that in order to conceive of myself as being mad, I need not suppose that my language is different from that of the people around me.

VI. *Insanity and the Public World*

This leads me to an aspect of the madness hypothesis very close to that raised by Frankfurt. There is a difficulty in supposing that I am now mad, and am in a position to give up my trust in my ability to judge about certain things in standard situations, and supposing at the same time that I use that very ability which I am supposing myself not to have to draw conclusions about the relationship of my thoughts and reality. That is self-defeating, for the conclusions are reliable only if the premise is false. Even if the conclusion is true this would only be so by chance.

Now it is true that even when I am mad the world I live in is a public world. Contrast the world which I dreamt about. Even if I dreamt that others made judgments about me, this would not mean that any of them were there near me in the world in which I was asleep and dreaming. The objects which I dreamt that I perceived cannot be perceived by other people, whereas, when I am mad, I am physically located in the same world as others whom I perceive and who are most probably observing or making judgments about me as well as about the things that I perceive. This is why in my thought-experiment of supposing myself to be mad the fact that their judgments and my own judgment are in disagreement is one part of what I have to account for. Even if a madman is not aware of the real disagreement between some of his beliefs and those of other people, he is often aware of the discrepancy between what others claim to believe and his own beliefs. (For this reason he may believe that others are lying or acting.)

There is no seeming disagreement now between how others claim to perceive things around me and how I perceive them. What can I conclude from this? Not much. For to conceive of myself as mad is, as we have seen, to give up my assumption about how reality may be expected to impinge on my beliefs, and the attitude of others is included in that reality. Obviously I notice that, even in my daily life which I call normal, my emotions or desires may make me ignore certain evidence or draw unwarranted conclusions. Expectations built up by my prejudices or past experiences may make me fail to perceive how things really are. But we assume (even if we have not always observed) that people, including myself, can be made to see how conclusions they have drawn do not follow, and correct or change their views by being exposed to further and further evidence. To suppose that I am mad is to suppose, among other things, my not having this ability.

VII. The Supposition That One Is Dreaming and the Supposition That One Is Mad

To suppose that I am merely dreaming is to suppose that I am in a state such that it is possible in principle to wake up

from it. I can understand what it is to wake up, and to recollect having dreamt, either by going through such an experience myself, or through learning about other people's accounts of such experiences.⁸

Professor Malcolm has said that one cannot make a judgment while one is sound asleep, because a deliberate action like making a judgment can only be done when one is fully conscious; indeed, as Wittgenstein remarks in *On Certainty*, section 383, to dream that one makes a judgment is not to make a judgment—any more than to dream that one proves a theorem is to prove a theorem, or to dream that one kills is to kill. And since any criterion people may give to distinguish real experiences from dream experiences can be *dreamt* to be satisfied, it will not help us. But it is far from clear whether to be aware that one is dreaming is merely to dream that one is aware. Can I not be aware that I am dreaming?

If one takes a Malcolmian position, it is impossible for a person to wonder whether he is dreaming and to wonder this while dreaming. Wondering is an activity which is intentionally carried out and thus cannot, according to this position, be carried out when a person is asleep; and a person's dreaming that he is wondering will not constitute wondering while he is asleep. I agree with this. But does this imply that the fact that I can wonder whether I am dreaming proves that I am *not* asleep and hence not dreaming? Or does it suggest that it makes no sense to conceive that I may be dreaming that I am wonder-

8. It is interesting to observe the importance that recollection plays in the views of dreaming embraced by thinkers otherwise very diverse. Normal Malcolm has asserted that it is only the telling of dreams, by people who recollect dreams, that has given, even to those who have never themselves recollected dreaming, the concept of dreaming. Michel Foucault, in a more empiricist vein perhaps, claims that recollection which we have when we have woken up gives us our concept of dreaming. The great difference between the two thinkers lies in Malcolm's thinking that it follows from *this* that one cannot sensibly suppose oneself to be dreaming now, whereas Foucault believes that it renders the supposition that one is dreaming a possible one. As a rational thinking subject, I can conceive of myself waking up to recall believing that I was having the kind of experiences I now have, and I can suppose myself judging that I dreamt that I was sitting before the table thinking about scepticism.

ing? No, neither. In the past I have sometimes dreamt that I was wondering whether I was awake and later woken up. Therefore, even though I accept the Wittgensteinian part of the Malcolmian position given above, it seems that the following points which diverge from what Malcolm holds are also true. (a) Dreaming that one perceives, like perceiving, is an event occurring at a specifiable time. (b) When I am awake it may always be possible in principle to judge that I am awake (though as a matter of fact I do not often do so). (c) Here we have the reason why I can seem to recall an earlier event and can wonder whether it was one of experiencing it or dreaming that I experienced the event.

I want to suggest that this view of dreaming also makes it possible to conceive in a *general* way, in respect of a time when I am not making a clear judgment about my being awake, that I may come after that time to realize that I was dreaming then. There is no need for me to take up a position as to whether I am dreaming that I am wondering or whether the wondering is an accompanying reflexive awareness and not itself a content of my dream. I believe that the conceivability of making a judgment that I have woken up at a later time is all that is required in order to conceive that I am dreaming. I need not conceive myself to be judging now in a dream that I am dreaming. As Professor Bernard Williams has said, it is not inconsistent to hold that, when one is awake, one can tell that one is, and also hold that, when one is dreaming, one cannot tell whether one is dreaming or not.⁹

Returning now to madness, it seems to me that the same applies. I can conceive, with respect to the present time *t*, that I may come to judge later that I was mad at *t*. But this is not to conceive now that I am mad. For the reasons I gave earlier, I cannot put myself in the position of *thinking* a mad person's thoughts. That does not, however, prevent me from envisaging that I am now in a state such that it is possible in principle that I shall in the future come to realize that I was mad. I can conceive now of myself realizing in the future that my beliefs and those of others had come adrift without my now being

9. Bernard Williams, *Descartes*, app. III, pp. 309-13.

aware of it. *There is no conceptual difficulty here, and in practice this has been the experience of those who recover from mental derangement. Such people recall having believed they were sane but are prepared to judge later that they were mistaken. In order to conceive that I may in the future come to mistrust my present thinking, I do not have to rely entirely on future recollections of my present experiences. What is at issue are future judgments about how the world once was, and about people's experiences of the past—judgments which will make me doubt the reliability of my present thoughts and beliefs.*

If one is to assess experiences or thoughts piecemeal, then of course there is nothing that makes later experiences or beliefs more certain just because they are later. We do not, however, assess the certitude or the reliability of experiences piecemeal. We make judgments about how the world was or how we must have been over a period of time, and then on the basis of this, we can deny the veracity of our past beliefs or of the description that we gave to ourselves of the objects of perception. It is because of this that I can conceive of my coming to judge that my present thoughts are unreliable, even if it makes no sense for me to doubt now that they are. (This is not to deny that I could be mad and remain so for the rest of my life. But for me to be mad something would have to be able to count as recovering from it.)

VIII. Normative Concepts

To allow myself to suppose that I am in a state such that something could count as recovering from it is, however, to acknowledge the distinction between mad and non-mad states. It is to acknowledge that there are states in which my judgment and others' judgments about the external world agree sufficiently, to enable me to accept a common framework without which we cannot even identify our mistakes and disagreements. This is to acknowledge of some of our states that they are normal states, i.e. that they are states such that, when we are in them, we can accept that reality is indeed impinging on our beliefs.

Some who believe that there is no fundamental difference

between the way madmen think and the way others think have suggested that the thoughts of all of us, mad and normal alike, are highly condensed, associative, symbolic, and overdetermined. This is probably true. But we still have to ask the question what thoughts our thoughts are supposed to be the condensations of. These are associations of what thoughts? And what thoughts do our thoughts symbolize? Isn't condensation the condensation of a complex of thoughts which the person believes to be true? Association leads from thoughts which one believes to be true to others which do not necessarily follow from them, but are linked with them in the person's mind so that the person ends up by holding the associated thoughts to be true. Symbolic thoughts are thoughts which symbolize for the person what he believes to be true. It seems to me that one cannot cut the fundamental link between beliefs and truth in the thinking which we take as the "norm". And in this normal thinking the awareness of the unsoundness of one's belief has an eroding effect on the belief itself.

For certain simple material objects, X's, when we who possess the concept X see an X in favourable perceptual situations, we see it as X. We do not understand without further explanation why others would not see it as X. When we see our hands we see them under normal conditions as hands. So it is not easy to understand what it is that people experience when they look at their hands with good eyesight, and do not see them as hands, or deny that they have a body.

I can conceive of myself seeming to recall having my arms amputated. I may also have reason to believe that somebody has given me an artificial arm. And I have to have some such story, a story that I could believe to be true, in order to dissipate the seeming incongruity between what I seem to perceive and what I believe. But the moment we make the story cohere with the available evidence, then we are no longer supposing ourselves to be mad. Either we are supposing ourselves to have a complex history such that what appears to us on first sight does not correspond to reality, or we are supposing ourselves to have an understandable false belief. (Thus, as Foucault points out, Descartes says only that if I were to deny that I had a body or deny

that my hands existed I would be *comparable* to someone insane. The comparison relates to what would be observable from the outside, and only concerns the likeness of oneself and of the madman in behaviour. He is not suggesting that one put oneself in the madman's position from the inside.)

Just as dreaming is contrasted to the waking state, madness is first of all understood negatively as something involving our not sharing the way a madman's mind operates in relation to the reality around him.¹⁰ (This is not the same as the claim that madness is a political category based on arbitrary discriminatory attitudes.) Jaspers was not being complacent when he defined the true delusions of the schizophrenics in terms of experience which could not be understood by others. Their beliefs and desires are held together in a system by processes we are not party to (where "we" means something like those of us who share thoughts together in the normal way). But in saying this we are not implying that we have a very clear view of what the normal is, or that there is one feature common to all those who are mentally ill. Nor are we implying that we believe in any accepted pattern of behaviour as having to be the normal one. (Even less then do we commit ourselves to any particular view about how we should *treat* the people we consider mentally ill.) But, if the idea of madness did not involve normative concepts of this kind, mad systems of belief would be no different from normal systems supported by beliefs with which we disagree. And to judge from the description given by specialists, as well as by those close to people who are thought to be mad, or even by the people themselves, it seems impossible not to admit that we do distinguish someone's thinking mad thoughts from his thinking false thoughts or having thoughts with which we disagree.

Madness is a loose concept. And, whatever madness or normality of thinking is, these are properties which admit of degree. I have suggested that the ascription of madness is based not so much on the person's having sensory experiences not

10. Even in the classical works of psychiatrists such as E. Bleuler and K. Schneider, or more recent works such as W. Winkler, the thinking of schizophrenics has been characterized negatively, e.g. loosening of associations, delusory logic, or non-syntactic thinking.

given to others as on his not seeming to resemble us in the manner in which individual pieces of perceptual evidence link up with whole sets of beliefs and desires. (For although psychotics are said to have auditory, olfactory, or even visual hallucinations, having hallucinations is not enough by itself to make a person be considered mad. People can hallucinate when they are extremely tired, when they have drunk too much, or when they have taken certain kinds of drugs, like mescaline. It is not even the case that people who hallucinate always take that which they hallucinate to be constituents of reality. And even when hallucinations produce false beliefs in me, when, for example, a drug I have taken makes it impossible for me to focus on objects at near distance, so that I simply cannot see my hands, and come to doubt whether I have hands, this is not irrational or mad.)

The men described by Descartes as mad are not merely people who have different beliefs from other people because their senses are somehow different from ours; they are people who seem to perceive certain phenomena *as* facts and events in a manner incomprehensible to us. The reason why they do this appears to us to be that their experience is incorporated into a system of beliefs that is impervious to what seems *to us* to be evidence. And we cannot understand how their beliefs can be impervious to reality in this manner. This is to acknowledge that any thought-experiment we can carry out is limited by our horizon of intelligible desires and patterns of thinking.

I have argued that the supposition that one is mad is possible although it carries with it a difficulty of a quite different order from the supposition that one's senses are deceiving one. No amount of checking by me of features of my experience or seeming certitude will establish anything about the relation of myself and reality once I enter into the supposition that I am now mad. But the possibility of my supposing that I am now insane and that among other things my cognitive powers are impaired when things appear quite normal to me, and when my recent experiences appear to me not to be odd, is based on the conceivability of my coming to judge retrospectively that I am mad. This presupposes my accepting a norm of sanity which

reposes on the ideas of how reality impinges on a belief, or how belief is shaped by reality. To say this is not to believe in the existence of raw "true" experiences that are free from theories or to say that there is only one system of concepts that truly "expresses" reality; *least of all is it to deny that often insanity has only been in the eyes of the beholder.*

his notion of rigid designation and the generation of, what I call, "exotic" necessary truths. I will try to show that the theory of natural kind terms must in fact use additional ideas and different argumentation. In doing this I will not, generally, be uncovering things only implicit in the writings of Kripke and Putnam; for the most part they are there to be seen. But I think it is easy to be misled about what is going on and, I think, the authors themselves sometimes go wrong in describing their own procedures. My own feeling is that when one sees more clearly what is involved some interesting puzzles emerge. At the end I try to develop an outline of one of these.

I

Before getting down to details it would be useful to look at some general features of the Kripke-Putnam treatment of natural kind terms. In the first place most of the examples they use are words and expressions in everyday use, such as 'water', 'tiger', 'gold', and 'heat'. While the theory calls for a certain relationship between the semantics of these terms and science, the terms obviously are not borrowed from the vocabulary of science and were part of English long before the advent of modern science. I think it is no accident that terms with these characteristics were chosen. In the first place, although one might suppose that if terms for natural kinds are to be found anywhere the language of science would be replete with them, it is not obvious that the Kripke-Putnam theory is applicable to kind terms in science. Nor is it obvious that it will apply to terms which the vernacular has borrowed from the language of science, such as 'plutonium' or 'electron'.

Putting aside, however, the doubts just expressed, there is a second reason for choosing as examples terms from the vernacular that antedate the rise of science: The Kripke-Putnam theory offers an answer to an important puzzle about the relationship of vernacular kind terms and scientific discovery. We seem willing to tailor the application of many of our vernacular terms for kinds to the results of science and if necessary to allow our usual means of determining the extension of these terms to be overridden. There is, for example, a product on the market com-

Kripke and Putnam on Natural Kind Terms*

KEITH S. DONNELLAN

The theory of natural kinds terms developed by Saul Kripke and Hilary Putnam is seen by both authors, I believe, as being intimately connected to Kripke's views about reference, perhaps even a consequence of them.¹ The views about reference, however, were first applied to singular terms and one wonders whether the transition to a theory about general terms, terms for kinds of things, can be made without more complications than either author, in my opinion, clearly indicates.

I want to examine, then, the application of some of the central ideas Kripke developed concerning singular terms to general terms for kinds. In particular, I will be concerned with

* Norman Malcolm has been a great influence on my philosophical thinking and remains so. I wish that my contribution to this volume in his honor were on a topic for which he is well known. A seminar on dreaming which I gave last year did not result in anything which I considered an original contribution on my part. Hence, this contribution of mine simply represents my own current philosophical interest. It benefits from discussion with, among others, Jay Atlas, Tyler Burge, David Kaplan, and Jon Wilwerding. The latter has independently developed an example somewhat like the one given by me in the last section, but from which he has been able to derive somewhat stronger conclusions than I do here.

1. For Kripke's views I use "Identity and Necessity" in M. K. Munitz, ed., *Identity and Individuation* (New York: New York University Press, 1971), pp. 135-64; and "Naming and Necessity" in Davidson and Harman, eds., *Semantics of Natural Language* (Dordrecht: D. Reidel, 1972), pp. 253-354. For Putnam's views I use "The Meaning of 'Meaning'" in his *Mind, Language and Reality* (Cambridge: Cambridge University Press, 1975), vol. 2, pp. 215-17.

his notion of rigid designation and the generation of, what I call, "exotic" necessary truths. I will try to show that the theory of natural kind terms must in fact use additional ideas and different argumentation. In doing this I will not, generally, be uncovering things only implicit in the writings of Kripke and Putnam; for the most part they are there to be seen. But I think it is easy to be misled about what is going on and, I think, the authors themselves sometimes go wrong in describing their own procedures. My own feeling is that when one sees more clearly what is involved some interesting puzzles emerge. At the end I try to develop an outline of one of these.

I

Before getting down to details it would be useful to look at some general features of the Kripke-Putnam treatment of natural kind terms. In the first place most of the examples they use are words and expressions in everyday use, such as 'water', 'tiger', 'gold', and 'heat'. While the theory calls for a certain *relationship between the semantics of these terms and science*, the terms obviously are not borrowed from the vocabulary of science and were part of English long before the advent of modern science. I think it is no accident that terms with these characteristics were chosen. In the first place, although one might suppose that if terms for natural kinds are to be found *anywhere the language of science would be replete with them*, it is not obvious that the Kripke-Putnam theory is applicable to kind terms in science. Nor is it obvious that it will apply to terms which the vernacular has borrowed from the language of science, such as 'plutonium' or 'electron'.

Putting aside, however, the doubts just expressed, there is a second reason for choosing as examples terms from the vernacular that antedate the rise of science: The Kripke-Putnam theory offers an answer to an important puzzle about the relationship of vernacular kind terms and scientific discovery. We seem willing to tailor the application of many of our vernacular terms for kinds to the results of science and if necessary to allow our usual means of determining the extension of these terms to be overridden. There is, for example, a product on the market com-

posed half of sodium chloride and half of potassium chloride. It looks like and tastes like ordinary salt. In most ordinary circumstances—in talking about how much to put in the stew, for example—we would be happy to call this product “salt” even if we knew its chemical composition. But if pressed to say whether this product is “really” salt, I think we would, if we know some elementary chemistry and the chemical composition of the product, concede that it is only half salt. To take a couple of more examples, I would give up calling a stone purchased as a diamond a ‘diamond’ if assured by experts that it did not possess a certain crystalline structure of carbon and I am prepared to be corrected when what I take to be a wolf in a cage at the zoo turns out to be identified by zoologists as being of a quite distinct species.

This apparent reliance on scientific results and classification raises a question about the extension and meaning of kind terms in the vernacular prior to scientific investigation. It is at least plausible that prior to modern chemistry, crystallography, and zoology a mixture of sodium and potassium chlorides, a sparkling clear crystal of some other substance than carbon, and an animal looking exactly like a wolf, but of that other species, might be called ‘salt’, ‘a diamond’, and ‘a wolf’ without a way for anyone to be the wiser.

Several questions come up at this point. Is the extension of the terms in question as used prior to the relevant scientific results the same as or different from the extension of the terms as used today? Is the meaning of the terms the same or different? And what is the status of the scientific results? To take the term ‘salt’ as an instance, it seems very plausible to say that at a certain point it was experimentally discovered that salt is the compound sodium chloride. Subsequently, knowledgeable speakers appear to use the property of being sodium chloride as the ultimate criterion for whether or not a bit of stuff is “really” salt. One common response to these observations is that prior to the scientific discovery being sodium chloride was not a part of what it meant to be salt, but that after the scientific discovery (for whatever reason), it became a part of (perhaps the whole of) the meaning of the term ‘salt’. This, however, has some unintuitive consequences. First, it implies that in this instance as in many

others a word has changed its meaning due to a scientific discovery. But we do not generally suppose that when we see a word such as 'salt' in, say, John Locke's *Essay Concerning the Human Understanding*, it has a different meaning. There is also the not unlikely possibility that if such a view of the situation is correct then the extension of such a term has also changed. And, finally, it is difficult to see how we can *now* speak of an experimental result if we have made it part of the very meaning of 'salt' that salt is sodium chloride.

The Kripke-Putnam view seems to have the virtue that none of these unintuitive results need be accepted. The facts as presented do not show, on their view, that, e.g., the word 'salt' has changed its meaning from, say, Locke's day to ours or that it has changed its extension. People in Locke's day might have called some stuff 'salt' that we would not (on the grounds that the stuff is not sodium chloride) but that does not show that the extension is different, only that (providing that our chemists have got things correct) people in Locke's day made mistakes which they had no means to correct. And finally that salt is sodium chloride was and remains a statement of an experimental discovery.

II

The way in which the results mentioned in the last section may seem to come about is through the application to kind terms of a now well-known, but still startling, result that Kripke obtained for proper names. Proper names, according to Kripke (and I agree with this) are what he calls "rigid designators". It will suffice for our purposes to characterize a rigid designator as a term which designates the same individual in all possible worlds. It will be recalled that in both "Naming and Necessity" and in "Identity and Necessity" Kripke introduces the notion first in connection with what are loosely called "singular terms". In particular, proper names are, in general, rigid while definite descriptions, for the most part, are not.

It will be recalled that one of the startling results Kripke obtains is that in regard to terms *rigidly designating contingently* existing objects (as opposed, say, to mathematical objects), certain sentences involving such terms turn out to express neces-

sary truths, although the fact that they express truths is to be learned by empirical means. I have called such truths "exotic necessary truths".²

I suppose one reason that this result is startling, aside from the fact that it represents a possibility not countenanced in philosophy before, is that since necessary truths are true in all possible worlds they do not distinguish the actual world from other possible worlds and so it looks as if an examination of how the actual world *is* could have no bearing on establishing the truth.

In any event, providing we accept the notion, rigid designators for individuals produce exotic necessary truths in a relatively simple manner. I stress this because I think matters to be different when we turn to terms for kinds.

Restricting ourselves to contingently existing individuals, exotic necessary truths arise when there are two rigid designators for the same individual. Kripke's main examples are sentences expressing the relation of identity in which, to put it loosely, the identity sign is flanked by two different rigid designators, designating the same individual. So if 'D' and 'D'' are two such terms the sentence

D is identical with D'

expresses an exotic necessary truth. And the argument is simply this: since by hypothesis 'D' and also 'D'' designate the same individual in all possible worlds they must designate that individual in each case. But since, by hypothesis, they designate the same individual in this world they must do so in all possible worlds. Hence, the sentence expresses a necessary truth. The sen-

2. I introduce the expression 'exotic necessary truths' not just to dramatize the interest of Kripke's discovery. The more obvious term '*a posteriori* truths' obscures an important point. If we distinguish a sentence from the proposition it expresses then the terms 'truth' and 'necessity' apply to the proposition expressed by a sentence, while the terms '*a priori*' and '*a posteriori*' are sentence relative. Given that it is true that Cicero is Tully (and whatever we need about what the relevant sentences express) 'Cicero is Cicero' and 'Cicero is Tully' express the same proposition. And the *proposition* is necessarily true. But looking at the proposition through the lens of the *sentence* 'Cicero is Cicero' the proposition can be seen *a priori* to be true, but through 'Cicero is Tully' one may need an *a posteriori* investigation.

tence expresses an exotic necessary truth for the following reason: While it may be possible for someone not to have to do any research about the actual world in order to know that such a sentence expresses a truth—as when, for example, one is introduced to a person and in the same breath given two names for him—it is obviously possible, and does occur, that a person or a culture has two names for the same individual without realizing it. So we have the stock examples of someone having both the names 'Cicero' and 'Tully' without realizing that, as they are used by him, they name one and the same person, and (what historically seems to me questionable) the story of the Babylonians having given the names 'Hesperus' and 'Phosphorus' to the planet Venus without realizing that a single heavenly body had been named. Under these circumstances, empirical investigation is required to ascertain that the identity sentence in question expresses a truth.

If this short exposition of Kripke's view about proper names is correct, it generates "exotic necessary truths" using essentially just the notions of rigid designation and identity.

What I want to argue is that this is not enough to give us the theory of natural kind terms that Kripke and Putnam want. Rigid designation (and identity), which gives the spectacular results for proper names, is not enough to do the same thing in the case of general terms for kinds. I do not mean to say that Kripke and Putnam would disagree with this, but Putnam says, for example, in his treatment of the term 'water', "Our discussion leans heavily on the work of Saul Kripke although conclusions were obtained independently" and then ". . . we may express Kripke's theory and mine by saying, for example, that the term 'water' is rigid."³

My view is that here Putnam is wrong about the mechanism of his own view. However, I also find in some of Kripke's descriptions of his argument the same mistake. Kripke does use an argument in the earlier version of his view, "Identity and Necessity," which does seem to extend in a simple manner the results and argumentation from proper names to (natural) kind terms. The problem is that it will not work.

But Putnam and Kripke (in the later work, "Naming and

3. "The Meaning of 'Meaning,'" pp. 230 and 231.

Necessity") use, on my reading, a more complicated argument. I want to question the latter argument as well, but to do so requires considerations which are more difficult to assess.

I want to argue that one cannot transfer, in a straightforward way, the Kripke results about proper names to terms for kinds, even though in the next section I assume that Kripke and Putnam must treat kind terms as being singular terms.

III

In trying to extend the notion of a rigid designator to non-singular terms we immediately come up against a problem which, I believe, neither Kripke nor Putnam touch upon. Rigid designation was initially defined by Kripke for singular terms—those terms which in some way carry along with them the requirement that there is a single individual referred to. A term such as 'tiger', unless it occurs in some such context as 'The tiger who bit my friend', does not at first glance seem to refer to any individual. Philosophers talk about the *extension* of a term and usually apply this notion both to singular terms, such as names, and general terms such as 'tiger'. The extension of a term is, roughly, what it applies to correctly.

Now if by talking about what is designated by a term in all possible worlds we were to think of the *extension* of the term 'tiger' we would not find it to be a rigid designator. In different possible worlds there are more or fewer tigers than in this and I suppose that the particular tigers of the actual world do not exist in all possible worlds.

For the purposes of this paper I am going to assume that construing terms for kinds, such as 'water', 'tiger', etc., as rigid designators *and* giving the Kripke-Putnam view the best run for its money is to think of them as what Mill calls "abstract" nouns. 'Tiger' is not to be thought of as designating its extension. Rather, it designates (is the name of) a certain species. 'Water' designates, not *its* extension—puddles, pools, ponds of stuff—but the substance, water. Thought of in this way, kind terms are in one way like proper names: they designate a single entity, albeit an abstract entity—a species or a substance in these

cases. I am aware that there are other suggestions about how to construe kind terms for the Kripke-Putnam theory. I don't believe that any are better suited, but the issue needs to be argued. All I will say here is that I think I can show that no other way has fewer problems for their view.⁴

If I am correct, a problem now arises. One would suppose that there is going to come out of the Kripke-Putnam view a distinction between general terms which designate natural kinds and general terms which do not. And there is at least the suggestion that rigid designation is the key. But, taken as "abstract" nouns, terms that seem intuitively to be obvious examples of non-natural kind terms seem just as rigid as those Kripke and Putnam use as examples of natural kind terms. Taken as abstract nouns there seems to be no difference at all between terms we would naturally suppose to fall in one class and terms we would naturally suppose to fall into the other in respect to rigidity. This was noted by David Kaplan in a footnote to his paper "Bob and Carol and Ted and Alice".⁵

One would naturally suppose that some such term as 'bachelorhood' would fall on the side of non-natural kind terms. But as an "abstract" noun it seems to be just as much a rigid designator as 'tiger' or 'water'. Rigid designation seems to provide no difference between natural and non-natural kind terms.

IV

Now, I want to identify the argument about "natural kind terms" which seems to me to be the straightforward extension of what Kripke shows about proper names and which seems to me not to work for general terms. I quote from a passage in Kripke's "Identity and Necessity" in which he seems to argue merely from the rigidity of the abstract nouns (or, to be accurate,

4. For one alternative to my assumption see Monte Cook, "If 'Cat' Is a Rigid Designator, What Does It Designate?" *Philosophical Studies*, 37, no. 1 (Jan. 1980). If Cook is correct much of what I say is wrong. I believe I can counter him, but to do so requires more than I can put into a footnote.

5. Hintikka, Moravcsik, and Suppes, eds., *Approaches to Natural Language*, p. 518 n.31.

Necessity") use, on my reading, a more complicated argument. I want to question the latter argument as well, but to do so requires considerations which are more difficult to assess.

I want to argue that one cannot transfer, in a straightforward way, the Kripke results about proper names to terms for kinds, even though in the next section I assume that Kripke and Putnam must treat kind terms as being singular terms.

III

In trying to extend the notion of a rigid designator to non-singular terms we immediately come up against a problem which, I believe, neither Kripke nor Putnam touch upon. Rigid designation was initially defined by Kripke for singular terms—those terms which in some way carry along with them the requirement that there is a single individual referred to. A term such as 'tiger', unless it occurs in some such context as 'The tiger who bit my friend', does not at first glance seem to refer to any individual. Philosophers talk about the *extension* of a term and usually apply this notion both to singular terms, such as names, and general terms such as 'tiger'. The extension of a term is, roughly, what it applies to correctly.

Now if by talking about what is designated by a term in all possible worlds we were to think of the *extension* of the term 'tiger' we would not find it to be a rigid designator. In different possible worlds there are more or fewer tigers than in this and I suppose that the particular tigers of the actual world do not exist in all possible worlds.

For the purposes of this paper I am going to assume that construing terms for kinds, such as 'water', 'tiger', etc., as rigid designators *and* giving the Kripke-Putnam view the best run for its money is to think of them as what Mill calls "abstract" nouns. 'Tiger' is not to be thought of as designating its extension. Rather, it designates (is the name of) a certain species. 'Water' designates, not *its* extension—puddles, pools, ponds of stuff—but the substance, water. Thought of in this way, kind terms are in one way like proper names: they designate a single entity, albeit an abstract entity—a species or a substance in these

cases. I am aware that there are other suggestions about how to construe kind terms for the Kripke-Putnam theory. I don't believe that any are better suited, but the issue needs to be argued. All I will say here is that I think I can show that no other way has fewer problems for their view.⁴

If I am correct, a problem now arises. One would suppose that there is going to come out of the Kripke-Putnam view a distinction between general terms which designate natural kinds and general terms which do not. And there is at least the suggestion that rigid designation is the key. But, taken as "abstract" nouns, terms that seem intuitively to be obvious examples of non-natural kind terms seem just as rigid as those Kripke and Putnam use as examples of natural kind terms. Taken as abstract nouns there seems to be no difference at all between terms we would naturally suppose to fall in one class and terms we would naturally suppose to fall into the other in respect to rigidity. This was noted by David Kaplan in a footnote to his paper "Bob and Carol and Ted and Alice".⁵

One would naturally suppose that some such term as 'bachelorhood' would fall on the side of non-natural kind terms. But as an "abstract" noun it seems to be just as much a rigid designator as 'tiger' or 'water'. Rigid designation seems to provide no difference between natural and non-natural kind terms.

IV

Now, I want to identify the argument about "natural kind terms" which seems to me to be the straightforward extension of what Kripke shows about proper names and which seems to me not to work for general terms. I quote from a passage in Kripke's "Identity and Necessity" in which he seems to argue merely from the rigidity of the abstract nouns (or, to be accurate,

4. For one alternative to my assumption see Monte Cook, "If 'Cat' Is a Rigid Designator, What Does It Designate?" *Philosophical Studies*, 37, no. 1 (Jan. 1980). If Cook is correct much of what I say is wrong. I believe I can counter him, but to do so requires more than I can put into a footnote.

5. Hintikka, Moravcsik, and Suppes, eds., *Approaches to Natural Language*, p. 518 n.31.

nouns and noun phrases) that a certain statement is an exotic necessary truth. Kripke has just been discussing the statement that heat is the motion of molecules. He says,

To state the view succinctly: we use both the terms 'heat' and 'the motion of molecules' as rigid designators for a certain external phenomenon. Since heat is in fact the motion of molecules, and the designators are rigid, by the argument I have given here, it is going to be necessary that heat is the motion of molecules.⁶

Some comments on this passage: First, Kripke takes the terms 'heat' and 'the motion of molecules' as functioning in the sentence 'Heat is the motion of molecules' as what I have called an abstract noun (or an abstract noun phrase). Second, he is treating the statement in question as an *identity* statement. Third, when he says "by the argument I have given here" I take him to mean the general argument of the paper as applied to this example. And the general argument started out with a consideration of proper names and includes the following passage:

If names are rigid designators, then there can be no question about identities being necessary, because 'a' and 'b' will be rigid designators of a certain man or thing *x*. Then even in every possible world, *a* and *b* [sic] will both refer to this same object *x*, and to no other, and so there will be no situation in which *a* might not have been *b*.⁷

So the argument seems to be the same one given in the case of proper names and to depend simply on the rigidity of the terms involved. And this may seem to go against my saying that it won't do to sum up the view about natural kind terms by saying that such-and-such a natural kind term is a rigid designator. For if the argument about proper names is sound, this one seems sound also. But there are problems. For example: I would suppose that both Putnam and Kripke would consider it likely that the following statements are, if true, necessarily true:

- (a) Tigers are mammals.
- (b) Water is a compound, not an element.

6. "Identity and Necessity," p. 160.

7. *Ibid.*, p. 154.

Neither of these, obviously, is an identity statement. So even given that the terms 'mammal' and 'a compound not an element' are *rigid designators of certain kinds*, the simple argument which can be applied to names cannot be used. Of course, if we already knew the truth of some identity statement from which (a) or (b) followed, and if the identity statement were an exotic necessary truth, then a simple extension of the argument would give the desired result. For instance, if we know that *heat is the motion of molecules* is an exotic necessary truth then we can simply deduce that *heat is some state or other of molecules*—the latter not being an identity statement.

But we need not know any such identity statement in order to discover (and historically I would guess this is the actual situation) that (a) or (b) is true. I am not even sure that in the case of (a), at least, we even now have any *identity* statement from which it can be deduced that tigers are mammals.

It might be said that all I have shown is that we can know that (a) and (b) are true without being in a position to know that they are exotic necessary truths: that to know the latter we would have to await the discovery of some exotic necessarily true identity statement. My point, however, will be that on the more complex theory, as I understand it, we *could* know that (a) and (b) are exotic necessary truths without awaiting the appearance of some identity statement which we can know to have that property. And this is another reason why we should not sum up Kripke's or Putnam's view by saying that this or that term is a rigid designator.

V

Now I want to imagine time-transporting John Locke to our era and convincing him of our scientific results and also teaching him what is meant by a rigid designator. I choose to transport *Locke* because I think his view about terms for natural kinds (he uses the expression "names of substances", but the topic is the same) represents an enemy for Kripke and Putnam. Locke, on my reading, is willing to allow that there might be something like natural *kinds* (divisions in nature as a consequence of the *primary qualities* of what we might now think of as such things

nouns and noun phrases) that a certain statement is an exotic necessary truth. Kripke has just been discussing the statement that heat is the motion of molecules. He says,

To state the view succinctly: we use both the terms 'heat' and 'the motion of molecules' as rigid designators for a certain external phenomenon. Since heat is in fact the motion of molecules, and the designators are rigid, by the argument I have given here, it is going to be *necessary* that heat is the motion of molecules.⁶

Some comments on this passage: First, Kripke takes the terms 'heat' and 'the motion of molecules' as functioning in the sentence 'Heat is the motion of molecules' as what I have called an abstract noun (or an abstract noun phrase). Second, he is treating the statement in question as an *identity* statement. Third, when he says "by the argument I have given here" I take him to mean the general argument of the paper as applied to this example. And the general argument started out with a consideration of proper names and includes the following passage:

If names are rigid designators, then there can be no question about identities being necessary, because 'a' and 'b' will be rigid designators of a certain man or thing *x*. Then even in every possible world, *a* and *b* [sic] will both refer to this same object *x*, and to no other, and so there will be no situation in which *a* might not have been *b*.⁷

So the argument seems to be the same one given in the case of proper names and to depend simply on the rigidity of the terms involved. And this may seem to go against my saying that it won't do to sum up the view about natural kind terms by saying that such-and-such a natural kind term is a rigid designator. For if the argument about proper names is sound, this one seems sound also. But there are problems. For example: I would suppose that both Putnam and Kripke would consider it likely that the following statements are, if true, necessarily true:

- (a) Tigers are mammals.
- (b) Water is a compound, not an element.

6. "Identity and Necessity," p. 160.

7. *Ibid.*, p. 154.

Neither of these, obviously, is an identity statement. So even given that the terms 'mammal' and 'a compound not an element' are rigid designators of certain kinds, the simple argument which can be applied to names cannot be used. Of course, if we already knew the truth of some identity statement from which (a) or (b) followed, and if the identity statement were an exotic necessary truth, then a simple extension of the argument would give the desired result. For instance, if we know that *heat is the motion of molecules* is an exotic necessary truth then we can simply deduce that heat is some state or other of molecules—the latter not being an identity statement.

But we need not know any such identity statement in order to discover (and historically I would guess this is the actual situation) that (a) or (b) is true. I am not even sure that in the case of (a), at least, we even now have any *identity* statement from which it can be deduced that tigers are mammals.

It might be said that all I have shown is that we can know that (a) and (b) are true without being in a position to know that they are exotic necessary truths: that to know the latter we would have to await the discovery of some exotic necessarily true identity statement. My point, however, will be that on the more complex theory, as I understand it, we *could* know that (a) and (b) are exotic necessary truths without awaiting the appearance of some identity statement which we can know to have that property. And this is another reason why we should not sum up Kripke's or Putnam's view by saying that this or that term is a rigid designator.

V

Now I want to imagine time-transporting John Locke to our era *and convincing him of our scientific results* and also teaching him what is meant by a rigid designator. I choose to transport *Locke* because I think his view about terms for natural kinds (he uses the expression "names of substances", but the topic is the same) represents an enemy for Kripke and Putnam. Locke, on my reading, is willing to allow that there might be something like natural *kinds* (divisions in nature as a consequence of the primary qualities of what we might now think of as such things

as atoms, molecules, genes, etc.), but that we don't have terms in the vernacular the extensions of which are determined by such divisions in nature. Instead, he believed that while men are inclined to believe that many of their kind terms are dependent upon nature in some such way, in fact the extensions of our vernacular kind terms are determined by the properties each of us uses to decide on the application of them. (To use Putnam's apt phrase, Locke holds that the meaning of kind terms are in "the head".) One might say that he held that while there may be natural *kinds*, there are no natural kind *terms*, at least in the vernacular.

But Locke transported here and now, given current philosophic and scientific theories, could, I believe, agree with both and still maintain his position. He could, that is, if the philosophic training only utilizes the argument based on rigid designation and identity that I find in the passages quoted from Kripke's earlier paper—the argument which seems to be the simple application of the ideas developed for proper names.

Using Kripke's example, he could admit that 'heat' and 'the motion of molecules' used as abstract noun and noun phrase are rigid designators and, given that he believes our science (or what we are pretending here is our science), that the extensions of these terms are the same. But he could not hold that any scientific discovery shows that the identity statement expressed by the sentence using these terms as *abstract nouns* is true.

Given that he would suppose, as I think he would, that we have different necessary and sufficient criteria for the correct application of the terms 'heat' and 'the motion of molecules' in our minds, there is no way for science to discover that these *kinds* of phenomena are identical; in fact, there is no way in which the *kinds* could be identical. The situation for Locke would be the same as it is in fact for the terms 'hearted thing' and 'livered thing'—as abstract noun phrases they designate different kinds, although their extensions are, in the actual world, identical.

What Locke easily concedes is that each kind term (or expression) names the very same *kind* in all possible worlds and so is a rigid designator. What he would deny is that we have discovered an *identity* between *kinds*. Perhaps we have discovered

a *co-extension* in the actual world, but that won't produce the desired result.

The reason why the argument we have looked at, using rigid designation and identity alone, works with proper names and not with kind terms is, of course, that proper names are singular and kind terms are general terms. If two *singular* terms are both rigid designators and their extension is the same, then there is just one individual that both terms designate in this, the actual, world and the same individual must be so designated in all possible worlds.

But, if we are dealing with two *kind* terms, their extensions might be the same in this, the actual, world without it being true that the two terms are names of *identical kinds*. What Locke might say, then, at this point is that *co-extension of instances* of two *kind* terms, even though each is a rigid designator as the name of a kind, doesn't show at all that the two *kinds* are identical.

VI

To obtain the same results for kind terms as Kripke obtained for proper names—especially the generation of “exotic” necessary truths—we need something more. And both Kripke and Putnam, as I read them, try to do this. I am going to use Putnam, but Kripke has the same augmented apparatus. Here are two connected quotes from Putnam about ‘water’:

To be water, for example, is to bear the relation same_L [same liquid] to certain things.⁸

And

x bears the relation same_L to y just in case (1) x and y are both liquids, and (2) x and y agree in important physical properties.⁹

These are two passages from the same paper in which Putnam says that his view and Kripke's can be summed up by saying that, e.g., ‘water’ is a rigid designator. But what new material is

8. “The Meaning of ‘Meaning,’” pp. 238–39.

9. *Ibid.*, p. 239.

introduced in these representative passages! Nothing about "important physical properties" is needed or used in Kripke's arguments about proper names to generate exotic necessary truths. Here then is the something more.

As a first shot I believe we can capture the idea in the following principle:

- (P) Something is water just in case it is a liquid and agrees in its important physical properties with the general run of stuff which English speakers call 'water'.

(That the term 'water' is mentioned does not, I believe, lead to difficulties as it would were we giving a classical definition.) The use of "the general run of stuff which English speakers call 'water' " can be explicated as follows: A word such as 'water' comes into a language as a word for a certain kind of stuff and ordinary users of the word will usually be able to apply the word by having learned to recognize certain characteristic "surface" properties—roughly those one can discern by relatively unsophisticated uses of the senses.

One complication I will ignore is that even before scientists, who are supposed to uncover the important physical properties, come into the picture there may be room for what Putnam calls a "division of linguistic labor" which is not reflected in (P). There may be users of the language who, although they have a term in their vocabulary and even may be said to "know the meaning", do not have a sufficient grasp of what surface properties to look for to be able to recognize instances of what the term is supposed to apply to. I think that this is probably what Putnam's situation is concerning the terms 'elm' and 'beech', a situation which I share with him. I cannot tell the difference between these two trees, but what I need—at least for such purposes as ordering from a tree farm—is not a scientist, but more likely a knowledgeable gardener. Since it would be absurd for anyone who wants to investigate the important physical properties of elms or beeches to consult me, even though I am an English-speaking user of the terms in question, a modification of (P) may be in order. But, as I say, I think we can ignore this here.

It does, however, seem in order to modify (P) so as to reflect

the important fact that Putnam wants his mechanism to function across possible worlds. To capture this, let us modify (P) to get:

- (P') For all worlds W , something is water in W just in case it is a liquid and agrees in its important physical properties with the general run of stuff which English speakers call 'water' in the actual world.

It follows from (P') that the extension of the term 'water' in any possible world is not determined by the surface properties which ordinary people may use to decide whether or not some bit of stuff is water, but rather by the important physical properties. And these, presumably, fall within the province of scientists. Now if being H_2O is one of the important physical properties of the general run of stuff English speakers call 'water', then that water is H_2O is what I have called an *exotic necessary truth*. Its truth is a *discovery of science* and by P' it is *true in all possible worlds*. But notice that this result is not a consequence simply of the fact that the term 'water', taken as an abstract noun, is a rigid designator, but rather from the more complicated features of (P'), in particular its use of the notion of *important physical properties* of samples of stuff in the actual world.

There may seem to be another interesting and important consequence of Putnam's mechanism. Since it is the important physical properties, properties to be discovered by what probably has to be a sophisticated science, and not the surface properties, which determine the extension of a term such as 'water', it may seem to follow that prior to the rise of such a science users of the term, relying solely on surface properties, can be wrong—at least in some cases—when even after a most careful examination they call some bit of stuff 'water'. For it may not be H_2O . This is a result Putnam in fact insists upon, and one I wish to raise a question about.

VII

In his article, "The Meaning of 'Meaning,'" Putnam makes use of an ingenious set of "Twin-Earth" examples. I want to do

something similar. I want to imagine two cultures which up to a certain point in their history are as alike as one can make them. In particular, the languages they speak are identical, or, at any rate, there is nothing up to the date mentioned which affords a basis for positing a difference. I want also to imagine that at a certain period in their histories each culture develops a sophisticated science and scientific view of the world and that these also are identical down to the smallest detail. Now let us take one of the terms for kinds which existed in the vernacular languages prior to the rise of science. I want to argue that although that term, by hypothesis, would exhibit no linguistic differences in the two languages prior to the rise of science, and although the sciences developed are identical, the term could come to have a different extension in each culture. Or, to be more accurate, I want to argue that Putnam's account of natural kind terms allows for this result.

It will be somewhat simpler for my purposes if I temporarily shift examples—changing water into gold. One of Kripke's examples of what I have been calling exotic necessary truths is the statement that gold has atomic number 79. I feel sure that Putnam would regard atomic number—the number of protons in the nucleus of the atom of an element—as an important physical property.

If I understand Putnam's mechanism correctly, he would also agree that John Locke might have been mistaken, say, about the composition of a ring he valued for its gold content if, in fact, the material it was made of did not consist mainly of atoms having atomic number 79, even if the foremost goldsmith of his day would have, on the basis of careful scrutiny of its surface properties, said it was undoubtedly gold.

For convenience we may as well imagine that my two cultures exist on Earth and Twin-Earth, much as in Putnam's examples. I will imagine then that my Twin-Earth has the same history as Earth up to some point in the earlier part of their twentieth century. In particular there has been a group of people, including the doppelgänger of John Locke, speaking a language called by them 'English' indistinguishable in every respect to an outside observer from the language of the same name spoken on Earth. At some point early in the twentieth century

they developed an atomic theory, once more indistinguishable from that developed on Earth. In particular they see the atom as having a nucleus composed of positively charged particles, which they call 'protons', and neutrally charged particles, which they call 'neutrons'. In their laboratories their scientists speak of 'elements' and use this term in the same way that our scientists have, *for those non-compound substances whose atoms have a particular number of protons in the nucleus—what they and we call the 'atomic number'.* They also recognize, and use the word, 'isotopes', isotopes being individuated by the combined number of protons and neutrons in the nucleus of the atom—that number being the isotope number.¹⁰ They also have, as our scientists have, come to realize that the same element may have, and usually will have, *several different isotopes and they give them names for use in the laboratory as our scientists have given the names 'protium', 'deuterium', and 'tritium' to the three known isotopes of what we call 'hydrogen'.* These similarities are all that I really need for the point I wish to make, so for the rest of the two developments in science, imagine that Twin-Earth and Earth are identical. I have thus, I believe, *so built my story that the scientific picture of the world is the same in the first parts of the twentieth century for both cultures.*

Although I believe that what I will now imagine is not strictly speaking necessary for what I will propose, it may help to make the situation psychologically more plausible. I imagine then that *on Twin-Earth not only do elements usually have several isotopes, but also it is a general rule that one of the isotopes of a particular element makes up the bulk of the element as it occurs in nature—the other isotopes being fairly rare.* I do not off-hand know whether this situation in fact obtains on Earth. If it does then I have made no change as yet *between Earth and Twin-Earth.* But even if it does not, the change I will have made is simply one concerning the relative distribution of things and will have nothing to do with a difference in scientific outlook,

10. For the sake of simplicity I am pretending that isotopes are distinguished by their isotope numbers alone. To accord with atomic theory as we know it I need a combination of atomic number and isotope number. This would complicate the telling of the Twin-Earth story, but the same point would emerge.

any more than the fact that we have less and less oil changes our scientific theories.

Now what isotope one is dealing with makes a big difference in how one will expect some bit of matter to behave. Just to mention a couple of the few things that I know about it, it is often the case that some isotopes of an element are radioactive and others not, and this, of course, has had profound consequences. Also some isotopes of the same element are unstable and break down into an isotope of another element. I believe this is the basis of carbon 14 dating. Of course, what element is in question is also of very great importance, especially in chemical reactions. But it might be a close question as to whether isotope number or atomic number has more importance.

Now with these suppositions and facts, it seems to me not psychologically implausible for my Twin-Earthlings to be more taken with, so to speak, the isotope number of a bit of substance rather than with its atomic number and also not implausible for them to diverge from our practice and to identify the substance designated by some of their vernacular natural kind terms not with a certain element, but with the isotope which makes up the bulk of what had been previously called by that term. Hence, for example, they identify gold, not with the element having atomic number 79, but with a certain isotope having a certain isotope number. The rare isotopes of the element having number 79 would then be dismissed as not "really" being gold, although, to be sure, in various ways very much like gold.

If I have described a possible Twin-Earth situation, then were my dopplegänger on Twin-Earth to ask someone in the Twin-Earth UCLA chemistry department what gold is, he would be told that it is a substance having such-and-such isotope number, while I, in the analogous situation, would be told that it is a substance having atomic number so-and-so. It should be obvious that if I and my dopplegänger defer to scientists as Putnam supposes we would, the extension of his term 'gold' and of mine will at this point diverge.

My story can, obviously, be extended to the vernacular term 'water'. On Twin-Earth, in my story, because isotopes are taken more seriously for one or another practical or historical reason,

we can suppose that Twin-Earthlings will identify water with protium oxide and exclude what we call 'heavy water'—deuterium or tritium oxide.

After the scientific discoveries and the mapping of non-scientific kind terms onto them, there will be a difference between Twin-Earthlings and Earthlings in regard to the truth-value of what certain sentences express. For example, there will be a sentence of the form 'Some gold has isotope number x and some gold has isotope number y ' which Earthlings will take to express a truth and Twin-Earthlings a falsehood. Correspondingly, while the sentence 'Gold has atomic number 79' will be regarded as expressing a truth by Twin-Earthlings provided 'has' does not carry with it the notion of identity, 'Gold is *identical* with the element having atomic number 79' will be regarded by them as expressing something false, while Earthlings will regard it as expressing something true.

We cannot immediately conclude that in my story it turns out that Twin-Earthlings regard certain *propositions* as true which Earthlings regard as false and *vice versa*, for it is open to one to say that the corresponding sentences do not express the same thing. In fact, in my story it would be paradoxical to conclude that the two parties have different beliefs about what is true or false because that would imply that there is a right and a wrong in the matter. And there does not seem to be any way of showing that either party is right while the other is wrong.

My talk about "mapping" ordinary language terms onto science may suggest a historical two-stage process: first there is science and its divisions of things into kinds, and then there is the hook-up of natural language terms—with some leeway in theory, if not in psychological reality, about how this second stage is carried out. I do not believe I need suppose this to be possible. Perhaps science has to begin with questions about kinds of things couched in ordinary language: "What is the nature of gold?", "What is the nature of water?", etc. Even if that is so, Twin-Earthlings in my story come up with verbally different things from Earthlings even though they start from what is to all appearances the same language and even though their science turns out to be identical. (Even if science must begin from questions couched in ordinary language, it soon develops its own

terms for natural kinds: The names of elements low on the periodic chart, in our science, are taken from ordinary language; those discovered later have invented names, and there are vastly more invented names for compounds than ones derived from ordinary language.)

What do I conclude from my story? I do not draw the conclusion that Putnam has failed to describe how natural kind terms in the vernacular function. The story does not show that. But I do think that two important points emerge. First, there is a certain slackness in the machinery which Putnam does not, I feel, prepare us for. In my story science develops apart on Earth and Twin-Earth. The slackness comes from how ordinary language terms for *kinds* are mapped onto the same scientific classifications. In my story I have envisaged only a small wobble; how much latitude there might be in theory I do not know. The point is that we can agree with Putnam's theory about how natural kinds terms function in ordinary language and still see that we might have done things differently even with the very same scientific results.

The second consequence flows out of the first. Putnam's view, it will be recalled, is that the extension of a term such as 'gold' or 'water' would be the same before scientific discoveries about important physical properties as it was afterward. John Locke would have been wrong, however justified he might have been, had he thought something to be gold or water which did not have the important physical properties discovered later. My strong inclination is to agree with Putnam that when I see the words 'gold' and 'water' in Locke's *Essay on the Human Understanding* there is no "change of meaning". And I also am strongly inclined to agree that I would not count something as, strictly speaking, *water* unless it were H_2O , or as *gold* unless it had the atomic number our scientists say that gold has. But I am an Earthling. In my story Twin-Earthlings do things slightly differently, although for no reasons having to do with a different linguistic basis or a different science. What should we then say about the extension of such ordinary language terms for kinds?

If you accept that my story of Earth and Twin-Earth contains no inconsistencies or other mistakes, then I do not see how we can accept Putnam's view that it is clear that natural kind

terms in ordinary language have the same extension before and after scientific discoveries and the mapping of those terms on to those discoveries. The "slackness" I have talked about seems to allow that from the very same linguistic base we may, after the very same scientific discoveries, move in different directions. Which way we move will change the extension. How can we then say that, given that we have moved in one direction, the users of the ordinary language terms might just have been dead wrong in applying the term to some particular instance? Locke's ring, which he took to be gold, might have been made of stuff having atomic number 79. In my story, the Earthling Putnam would say that Locke was correct. But suppose his ring were made of one of the rare isotopes, as we would say, of gold. Twin-Earthlings would say the ring wasn't made of gold at all. Putnam's doppelgänger would conclude that Locke was dead wrong. But nothing about language or science explains this difference.

You might object at this point that I have not kept my Twin-Earthlings as much the same as us as I have made it seem. After all, they do take a different turn; and must this not mean that they had at least one psychological difference from us?—they had the disposition to be struck more by isotopes than by elements. And perhaps this disposition in some way ought to count as showing a difference in the language, however much it looked to be the same as ours prior to scientific discoveries. Even if we admit that there was at some point a disposition on their part which we do not have, the disposition may not always have been there. We need not suppose that because the twentieth-century Twin-Earthlings had this disposition, the eighteenth-century ones did also. Nor do we need to suppose that there was such a disposition even in twentieth-century Twin-Earthlings prior to their looking at their scientific theory. Thus we seem to have no reason to think that their John Locke and ours differed psychologically or linguistically. Or rather we have none unless we accept what seems to be an outrageously bizarre view of language—that the extension of one's terms may be determined by the psychological quirks of some people several centuries hence.

There is one point about the scientific situation in regard to atomic theory that I have left out of my story. Historically, the discovery of the possibility of isotopes, and of the big difference

terms for natural kinds: The names of elements low on the periodic chart, in our science, are taken from ordinary language; those discovered later have invented names, and there are vastly more invented names for compounds than ones derived from ordinary language.)

What do I conclude from my story? I do not draw the conclusion that Putnam has failed to describe how natural kind terms in the vernacular function. The story does not show that. But I do think that two important points emerge. First, there is a certain slackness in the machinery which Putnam does not, I feel, prepare us for. In my story science develops apart on Earth and Twin-Earth. The slackness comes from how ordinary language terms for *kinds* are mapped onto the same scientific classifications. In my story I have envisaged only a small wobble; how much latitude there might be in theory I do not know. The point is that we can agree with Putnam's theory about how natural kinds terms function in ordinary language and still see that we might have done things differently even with the very same scientific results.

The second consequence flows out of the first. Putnam's view, it will be recalled, is that the extension of a term such as 'gold' or 'water' would be the same before scientific discoveries about important physical properties as it was afterward. John Locke would have been wrong, however justified he might have been, had he thought something to be gold or water which did not have the important physical properties discovered later. My strong inclination is to agree with Putnam that when I see the words 'gold' and 'water' in Locke's *Essay on the Human Understanding* there is no "change of meaning". And I also am strongly inclined to agree that I would not count something as, strictly speaking, *water* unless it were H_2O , or as *gold* unless it had the atomic number our scientists say that gold has. But I am an Earthling. In my story Twin-Earthlings do things slightly differently, although for no reasons having to do with a different linguistic basis or a different science. What should we then say about the extension of such ordinary language terms for kinds?

If you accept that my story of Earth and Twin-Earth contains no inconsistencies or other mistakes, then I do not see how we can accept Putnam's view that it is clear that natural kind

terms in ordinary language have the same extension before and after scientific discoveries and the mapping of those terms on to those discoveries. The "slackness" I have talked about seems to allow that from the very same linguistic base we may, after the very same scientific discoveries, move in different directions. Which way we move will change the extension. How can we then say that, given that we have moved in one direction, the users of the ordinary language terms might just have been dead wrong in applying the term to some particular instance? Locke's ring, which he took to be gold, might have been made of stuff having atomic number 79. In my story, the Earthling Putnam would say that Locke was correct. But suppose his ring were made of one of the rare isotopes, as we would say, of gold. Twin-Earthlings would say the ring wasn't made of gold at all. Putnam's doppelgänger would conclude that Locke was dead wrong. But nothing about language or science explains this difference.

You might object at this point that I have not kept my Twin-Earthlings as much the same as us as I have made it seem. After all, they do take a different turn; and must this not mean that they had at least one psychological difference from us?—they had the disposition to be struck more by isotopes than by elements. And perhaps this disposition in some way ought to count as showing a difference in the language, however much it looked to be the same as ours prior to scientific discoveries. Even if we admit that there was at some point a disposition on their part which we do not have, the disposition may not always have been there. We need not suppose that because the twentieth-century Twin-Earthlings had this disposition, the eighteenth-century ones did also. Nor do we need to suppose that there was such a disposition even in twentieth-century Twin-Earthlings prior to their looking at their scientific theory. Thus we seem to have no reason to think that their John Locke and ours differed psychologically or linguistically. Or rather we have none unless we accept what seems to be an outrageously bizarre view of language—that the extension of one's terms may be determined by the psychological quirks of some people several centuries hence.

There is one point about the scientific situation in regard to atomic theory that I have left out of my story. Historically, the discovery of the possibility of isotopes, and of the big difference

it makes which one is involved, came after the development of the simpler atomic theory which included the notion of elements and atomic numbers. Possibly this historical accident had an influence on the way in which we mapped our vernacular natural kind terms onto science. But this would be of little comfort to Putnam, I should think. It seems just as bizarre to suppose that the extension of one's terms should depend upon such future historical accidents as it does to suppose that they depend upon future psychological quirks.

To sum up then, if we go along with Putnam a certain distance we seem either to have to embrace unacceptable views about language or to admit that nature, after all, does not fully determine the extension of vernacular natural kind terms, and science is not wholly responsible for discovering their true extensions.

Discovering Essence¹

JOHN V. CANFIELD

A name giver, emulating Adam, attaches a kind term to a kind. Someone picked out this tiger and that and laid it down that they and others of this sort are to be called tigers. Thus was the reference of "tiger" established. On Saul Kripke's realist theory of kind terms, the rest is up to science.² Since the term was introduced by pointing to things in the world, we can discover their essence empirically. Science will do this by uncovering the inherent makeup of the things or stuff that someone in our cultural past named. In another of Kripke's examples, it may examine certain samples and discover that the essence of water is to be H₂O.

For such theories, items belong to the extension of a natural kind term in virtue of standing in a certain relationship to the term. Theory will spell out what the relationship is, and thereby say what makes it true that an item belongs to the term's extension.³ For Kripke, the relationship is that items in the extension

1. This paper has been revised in response to detailed criticisms the editors made of an earlier draft; I am very grateful to them.

2. "Naming and Necessity," in D. Davidson and G. Harman, eds., *Semantics of Natural Language* (Reidel: Dordrecht, 1972), pp. 253-355. Page references in the text are to this volume.

3. I believe Kripke seeks such a theory, despite his insistence that in the case of proper names he provides only a "picture" and not a theory, and despite his related raising of the question of whether analysis is possible at all. I

have the properties necessary for belonging to the same kind as the things to which the term was originally attached.⁴ This lump of metal is gold because it shares with the items originally labelled "gold" the scientifically discovered, essential property of being an element with the atomic number 79.

One rival theory is a form of linguistic relativism. Essence is not *de re* but *de dicto*. Arbitrarily set out criteria determine what belongs to a natural kind. The realist rejects the appeal to criteria. For him, if something is H₂O it is water, even if it looks, feels, and tastes unlike what we, by present or past criteria, would call "water". Reality itself and not criterial rules of language determines essence.

The relativist alternative has lately been associated with an epistemic interpretation of Wittgenstein's idea of a criterion.⁵ That interpretation is mistaken; the exegetically and substantively correct idea of a criterion is the one sketched in Norman Malcolm's review of the *Philosophical Investigations*.⁶ I have developed and supported his reading of "criterion" elsewhere.⁷ Here I apply it to the question of discovering essence, and the discussion of Kripke's theory of kind terms. I will establish that neither Kripke's arguments for his position nor those against its alternatives are conclusive, and I will bring out certain defects in his theory. I will then show how Wittgenstein's view of criteria yields a viable account of kind terms, and the scientific discovery of essence.

think a "picture" as he understands it is a stage in the as yet uncompleted attempt to get a theory. We get the theory by finding a way to remove the circular elements present in the statement of the "picture."

4. As Shoemaker put it in comments on an earlier version of this paper.

5. I am thinking of Gordon Baker and P. M. S. Hacker, who acknowledge their indebtedness to Michael Dummett, and who are also heavily influenced by Sydney Shoemaker's remarks on criteria in *Self-Knowledge and Self-Identity* (Cornell University Press: Ithaca, N.Y., 1963). See P. M. S. Hacker, *Insight and Illusion* (Oxford University Press: Oxford, 1972); Michael Dummett, "Truth," *Proceedings of the Aristotelian Society*, 69 (1958-59): pp. 141-62; and John T. E. Richardson, *The Grammar of Justification* (St. Martin: London, 1976).

6. Norman Malcolm, *Knowledge and Certainty* (Prentice Hall: Englewood Cliffs, N.J., 1963), pp. 96-129.

7. In *Wittgenstein: Language and World* (University of Massachusetts Press: Amherst, 1981).

I

First I will state Kripke's position more fully, focusing on points that are obscure or incomplete.

"One might," he says, "... discover essence empirically" (322). We can suppose, for example, that "scientists have investigated the nature of gold and have found out that it is part of the very nature of this substance, so to speak, that it have the atomic number 79" (318). Again, "Science can discover empirically that certain properties are necessary of cows, or of tigers" (322-23). Science discovers essence.

But does the scientist merely discover that gold has the atomic number, thereby discovering essence; or does he, alternatively, discover *that* it is the essence of gold to have this number? Columbus discovered a new land; but since he believed the place he sailed to was the Indies, we would not say he discovered that this place was a new land. *Discoveries that* I will call intentional, and the others extensional.

If for Kripke the scientist's discoveries of essence are merely extensional, then apparently a crucial element of a theory of essence is missing. For first, not all features discovered by science are essential ones. Gold has been discovered to have the following properties: It is a face-centered cubic metal. It has a lattice constant $a = 4.0699 \text{ \AA}$ at 20°C . Its melting point is $1,063^\circ\text{C}$, and its boiling point $2,600^\circ\text{C}$. It has a coefficient of linear expansion of 14.2×10^{-6} at 20°C . It has a tensile strength of about 7.5 tons and a Brinell hardness of 25. (From *Chambers Encyclopaedia*.) It is implausible that all these features are essential. Gold might, logically, have the other properties, but a very slightly lower coefficient of expansion than it now possesses. Since not all discovered features are essential, a theory of essence is presupposed in passing from the proposition that water is H_2O to the proposition that it is of the essence of water to be H_2O . But what theory?

The text, however, indicates that for Kripke the scientist's discoveries of essence are intentional; he discovers, say, that it is of the essence of water to be H_2O .⁸ This seems an odd thing

8. See for example the above quote about cows and tigers.

for a scientist to discover. How would he convince his colleagues that this is an essential property? By showing that there are no logically possible counterexamples?

If the discoveries are intentional then the scientist employs, explicitly or implicitly, a rule or standard for picking out essential from inessential features. If we knew that rule then we would have a theory of essence.

But it is obviously wrong to say that the scientist discovers *that* certain properties are essential ones. The way to read Kripke here, though it goes against the text somewhat, is as follows. There is a certain class K of features. In discovering that a kind has one of these features, the scientist discovers essence extensionally. As soon as, and in so far as, he discovers that a kind has such a feature, he discovers the essence of the kind. The governing rule is simply that all features in K are essential. It is the philosopher who establishes independently that all such features are essential ones. This is established, presumably, by certain arguments given in *Naming and Necessity*. Kripke owes an account of how to pick out the features that belong in K. To carry on our examination, we can pretend we understand how to extend the class of features given in his examples to new cases; and to mark that pretense I will call the features *k*-features.

A further question is this: What sample does a scientist examine in his search for *k*-features, and how is it related to what Kripke calls the original set, that is, the sample picked out in the (assumed) original fixing of reference of the kind term, or involved in the, as it were, original baptism of the kind.

It is obviously false that the scientist typically examines items from the original set itself. These are lost in the mists of time. The animals to which the term "tiger" was first attached are long since dead and dust; and it would be a joke to give someone the task of tracking down the specific volumes of fluid used in first fixing the reference of the word "water"; and so on.⁹

9. See the discussion by Eddie M. Zemach in "Putnam's Theory on the Reference of Substance Terms," *Journal of Philosophy*, 73, no. 5 (1976) (especially pages 123, 124) where this point is made and criticisms very similar to the ones developed below are given.

In response to the practical unavailability of items in the original set Kripke must say the following. The reference of "gold" is fixed by means of items in the original set. This set is extended in a certain way. Items are added to it if they have properties that the language users consider characteristic of the items in the original set. "Certain properties, believed to be at least roughly characteristic of the kind and believed to apply to the original sample, are used to place new items, outside the original sample" (320). It is items in the thus extended set that science will study to discover the essence of gold, and so on.

The theory has great initial plausibility. The reason is that it illegitimately borrows the plausibility of a certain historical truism. When scientists analyze and study something like gold, they begin with samples of it; and the samples are related in something like the way indicated above to the items used when the word "gold" was attached by convention to a natural kind. But Kripke goes beyond this truism; he attempts to turn plain facts about kind terms into a philosophical theory of them.

His theory maintains that *necessarily* it is of the essence of gold to be (say) an element with the atomic number 79 if and only if the items in the original sample, or most of them, are elements of this number. This *if and only if* statement does not follow from the empirical claim that our calling this and this "gold" has a history of the kind in question. One may not wish to question the Kripkean sketch of the history of the use of "gold", but it is quite another thing to grant the truth of the theoretical statement. The *only if* part of it is particularly vulnerable, as I will show.

There was a petty ruler in a corner of England in the Dark Ages who had a penchant for naming things and who enforced his terminology on people. He had heard about a metal, shiny, beautiful in color, malleable, and highly durable. For the pleasure of naming it, he sent for some. His councilors, however, played a joke on him. They prepared a package of coins, jewelry, and plate constructed from fool's gold, and pretended that these things were newly arrived samples of the metal to be named. The king duly fixed the reference of "gold" in terms of the items in the package. The councilors then took away the package, labelled it "original samples for 'gold'" and hid it.

under a stone in one of the castle dungeons. The extended set was extended from the original sample in terms of "certain properties, believed to be at least roughly characteristic of the kind and believed to apply to the original sample," namely the ones listed above, and others. And this is the actual history of our use of our term "gold", for the use spread from that kingdom to its present place in English. The councilors left behind a detailed record of their trick, with many proofs of authenticity. When these records are discovered in modern times, their legitimacy verified, the original sample uncovered and analyzed, will we say that all the gold bullion in the vaults of the world, all the gold watches, gold wedding rings, and coins, and so on, are not gold? Will we say that we have discovered that it is false that gold has the atomic number 79? Will we, in short, accept and apply the necessary condition laid down by Kripke's theory? Obviously not. The reason is that what counts as gold is in part a function of present and past conventions of language. Jewelers, metallurgists and others are taught what gold is in terms of paradigmatic instances and standard tests. We will not and should not change what we count as paradigmatic instances of gold in the face of the discovery of that particular original sample. We will simply note that someone played a strange joke on the king, and that as a result our use of "gold" got off to a peculiar start.

When we focus on the question of how the original sample is related to the one the scientist actually examines, we see, then, that the theory, at least in its present form, falls to counterexamples. And it may not be as easy to amend as one might at first suppose. For example someone might offer the following alleged minor modification of the theory. All the example shows, it might be said, is that the operative notion is to be not "original set" or "original sample" but something like "currently agreed standard samples." But this modification is not minor. Originally, the theory has an account of why the stuff the scientist examines is gold, in terms of its being in a certain relationship to items in the original set. But the proposed modification starts with the datum that this stuff he is to examine is gold, and thus it leaves unexplained something the theory wants to explain.

II

One of Kripke's arguments for *k*-features being essential is, apparently, of the *either or* type. Either the rival to his view is correct, or his is; the rival is wrong, therefore. . . . This reading is of necessity impressionistic, but I think the argument occurs in Kripke's paper, and I will now examine it.

In criticizing Searle, Kripke painstakingly spelled out the nature of the "associated descriptions" view of names; but he fails to demarcate carefully the target of attack with respect to kind terms. This imprecision vitiates his *either or* argument. In particular, he does not distinguish four different things that fall under the heading of identifying features.

These are as follows. (i) Sometimes, and this is the paramount understanding, identifying features are those used to pick out the members of the original set. I will call these "initial features". These are cited in fixing the reference of the term. Kripke sometimes assumes also that (ii) identifying features are those used *now* to pick out the items that belong to a kind. But there is no reason why the initial features must be the same as those used later. The "cluster concept" interpretation that Kripke criticizes is associated among other things with the idea of (iii) a dictionary definition of a term.¹⁰ "Cluster concept" is also associated with the idea of (iv) a criterion.¹¹ Kripke too uses the term "criteria" in the context of his attack on the identifying features view, indicating that he thinks of criteria as identifying features.

There is no reason for the features cited in an initial reference fixing to be entered in a dictionary definition of the term, nor as part of the criterion governing the term. (It is well to

10. See especially Kripke's discussion centering around page 318 of "Naming and Necessity."

11. The four types I have isolated show up, e.g., in this quote from page 318: "We . . . might . . . find out tigers had none of the properties by which we originally identified them. Perhaps none are quadrupedal, none tawny yellow, none carnivorous, and so on; all these properties turn out to be based on optical illusions or other errors. . . . So the term 'tiger' does not mark out a 'cluster concept' in which most, but perhaps not all, of the properties used to identify the kind must be satisfied". (The properties discussed were earlier cited as part of a dictionary definition of "tiger.")

remember, also, that both criteria and dictionary definitions can change.)

Because Kripke fails to make these distinctions, some of his arguments commit a fallacy of ambiguity, providing we take them to have the interesting conclusion that a cluster concept view of kind terms is, in all its plausible variations, wrong.

Kripke argues that something could fail to be of a certain kind even though it has all the identifying features of the kind:

There might be a substance which has all the identifying marks we commonly attributed to identify [*sic*] the substance of gold in the first place, but which is not the same kind of thing, which is not the same substance. We would say of such a thing that though it has all the appearances we initially used to identify gold, it is not gold. Such a thing is, for example . . . iron pyrites or fool's gold. . . . We can say this not because we have changed the *meaning* of the term gold, and thrown in some other criteria which distinguished gold from pyrites (316).

This example establishes the contingency of the connection between features and kinds for only two of the four types of identifying features. If these are epistemic—either (i) or (ii)—then the point holds. But in fact some dictionary definitions of gold are not in terms of features that gold shares with iron pyrites. (The O.E.D.: "Gold. The most precious metal. . . ." ¹² Secondly, it would be feckless to state a criterion for gold in terms of the features gold and iron pyrites have in common. On any reasonable understanding of "criterion" the criterion for gold differs from that for iron pyrites. It may seem on a quick reading as if this example is relevant to identifying features *qua* criteria and *qua* meanings (= dictionary definitions); but

12. Of course Kripke may well mean by "dictionary definition" something like "correct dictionary definition" and may in addition understand a correct definition as excluding such phrases as the one quoted from the O.E.D., on the ground that this cites an incidental fact about gold and not one of its (truly) defining features. If this is so, then his claims about "dictionary definition" amount to less than they seem; he is not really talking about dictionary definitions as they exist, warts and all, but about some idealized notion of a dictionary definition. On this understanding it is plausible to think that "correct dictionary definitions" amount to no more or less than clusters of criterial properties. But it is still very much in point to distinguish criteria (and hence also "dictionary definitions") from initial features.

in fact it is not. It establishes nothing about identifying features or cluster concepts in general but only something about identifying features *qua* (i) or (ii).

Could tigers be reptiles? An example can be constructed in which: (a) "tiger" is applied to a kind of animal on the basis of certain readily perceptible features; (b) these animals are found to have certain further properties, for example being mammals; and (c) a further kind of animal is found that, although it has the initial features in (a), is distinguished from tigers because it lacks the discovered features. Thus a people might refuse to call these creatures "tigers" although they are outwardly tigers; for none of the new kind has the mammalian features of bearing their young live, or suckling them.

Such an example, however, does not count against a cluster definition view of kind terms, unless one assumes, as one should not, that defenders of these views are committed to excluding the discovered features in question from the cluster, or from the criterion. On the contrary, a defender will say that the criterion governing the present use of a term may well differ from any criteria that governed it in the past. If it is the present use of "tiger" or "gold" or whatever that we are interested in, then what is in question is the criterion that currently governs the term. I am not modifying the cluster or criterial view here, in order to save it from the kind of example Kripke gives. The example can only be relevant to Kripke's conclusion if one falsely commits a cluster or criterial theory to the thesis that the initial properties are all and only those to be included in any cluster definition or in any criterion. This restriction is false to the letter and spirit of Wittgenstein's view, which emphasizes that the criteria governing a term may change. For example Wittgenstein notes the possibility of a fluctuation between symptoms and criteria (see *Philosophical Investigations*, sec. 79).¹³

13. At this point worries about change of sense may arise. It may be felt that if we allow the criterion governing a term to change over time, then this implies a change in sense; and against this it may be argued that it is an advantage of the Kripke-Putnam view that it repudiates such alleged changes of sense. I deal with this objection below, the basic reply being that a critical view is not committed to holding that, on some unacceptable sense of "change

III

The thesis that *k*-features are essential is supported by a second line of argument that sometimes accompanies and sometimes supplements the *either or* strain of argument.

The logical structure of this is as follows. We construct a possible state of affairs, and ask ourselves what we would say about it. Would we say, in this or that imagined case, that these things are tigers, or this stuff gold? The answers we get go against a criterial view, it is claimed, for we have imagined a case where the criterial features are absent and where yet we say that it is of the kind in question, or a case where the criterial features are present and we deny that it is of the kind. At the same time the answers—the result of “intuition”—are held to be in conformity with Kripke’s theory.

This argument from intuition is inconclusive for two reasons. First, similar examples just as strongly yield intuitions in conformity with the criterial view and opposed to Kripke’s theory. Second, the criterial view is embedded in Wittgenstein’s philosophy as a whole, which provides a certain way of interpreting appeals to “intuition”. For Wittgenstein, we should focus not on what we have an urge to say, but on the meaning of what we feel impelled to say. On his conception, the results of intuition here really only reflect hidden stipulations. These judgments of intuition presuppose that we extend the normal use of kind terms to cover the *outré* and abnormal cases imagined. These extensions in turn presuppose a choice of one or another criterion governing the kind term in the judgment made by intuition. There are alternative criteria that one might legitimately presuppose, so that the results of the intuitive judgments do not support Kripke over Wittgenstein, here, nor vice versa.

Let us look at these points in terms of some examples.

According to Kripke, “We might . . . find out tigers had *none* of the properties by which we originally identified them. Perhaps *none* are quadrupedal, *none* tawny yellow, *none* car-

of sense”—for example the one Putnam argued against—such a change has occurred.

nivorous, and so on; all these properties turn out to be based on optical illusions or other errors, as in the case of gold" (318). But let us fill in some details in this example and look more carefully.

Some explorers see animals, in fact jungle fowl, and dub what they see "tigers". Due, however, to the optical illusion in operation they perceive these chickens to have all the features we associate with tigers. The illusion is consistent, and has auditory, tactile, and olfactory dimensions as well. Everyone who later has contact with chickens of this kind sees them as tigers, sees their motions as tiger-like, hears their cluckings as roars, and so on. Certain chickens are caught, and exhibited, and their apparent size and ferocity inspire awe. One day the air clears of the illusion-causing particles, and "tigers" appear in their true shape, size, and behavior. Would we say that there are no tigers, or that these fowl after all are tigers, but that we had a misimpression of them?

If someone said in these circumstances, "It turns out that there are no tigers", I should certainly understand him, and grant the truth of his claim. On the other hand if someone said in the same situation that there are tigers, and that this creature in front of us is one, I would understand and grant that too, if I knew the facts stated above.

I am not supposing here that the explorers already possess the word "tiger" when they meet these creatures, and that this word they possess is our word. What I am imagining rather is that our use of "tiger" was introduced by the explorers, and that our use was sustained through the years by the illusion. In this case I do not see that one is forced to say that these things are tigers, but that tigers are different from what we thought. One could say this. But one could also say that it turns out that there are no tigers.

The person who denies there are tigers might have expressed himself more fully by saying: "The tigers of the jungle that were feared as man-eaters, that Blake wrote about, and so on, do not exist; it was all a quite incredible illusion." The other person's remark may be read along these lines: "When we were in the presence of the animals we have called tigers, and thought that we sensed tawny yellow, black striped, four-legged beasts,

and so on, it turns out that what was before us was a creature of totally different properties. We are calling these chickens 'tigers' since it was them we falsely perceived, and that we originally named 'tigers'".

Behind both claims are implicit stipulations. In the first it is supposed that something is a tiger only if it has at least some of certain features. In the second it is supposed that something is a tiger if it is the object pointed to in a certain reference-fixing act, or if it lies behind, in a certain way, our prior misperceptions.

It is false that only the second of the above responses is correct. Nothing stops us from using the term "tiger" in a way that implies that chickens, because they lack certain properties, are not tigers, whether or not these chickens were somehow causally involved in the dubbing of tigers. Indeed, to say that the chickens are not tigers seems a natural response to the discoveries imagined. Kripke can appeal here to what we *might* say in response to the uncovering of the illusion. But it is a mistake to think that, just because we might say this, the criterial view is wrong.

But let us look at the question of whether Kripke can, by appeal to intuition, show that only the judgments that are in line with his theory are correct. What must be at issue, in our example, is whether on the normal or ordinary or natural language use of "tiger" we would, in the imagined case, say that these things are tigers. This question is mistaken. The word "tiger" is used in certain normal circumstances. Our case is extraordinarily abnormal. We imagine the perceptions normally keyed to the use of "tiger" to be illusory; and then we ask what we would say in these circumstances. It is no good appealing to intuitions about the use of language here, for we are beyond the range of the normal use of "tiger". Our normal rules do not apply in this case. If we are to use "tiger" now we must, explicitly or implicitly, extend the rules to cover this new circumstance. We can say either that they are tigers, or that they are not. Whatever we say, we extend our earlier way of speaking to this abnormal case. There are different ways we could extend that normal way of speaking, and no one way is right to the ex-

clusion of the other (although Kripke attempts, by various arguments not yet examined, to prove the contrary).

These examples differ from one developed at the end of Section I, because there our judgment that this stuff is gold is based on our normal method of judging such things, applied in normal circumstances. The difference imagined is that in the mists of history the term "gold" had its reference fixed in a peculiar way. In asking whether we would apply the term "gold" to this or that stuff, in everyday life, our answer need not take into consideration any facts about the ancient history of the term, whereas in these cases of Kripke's we are to imagine that the ordinary means we have for making such judgments—the features of the stuff that we see, feel, measure, and so on—are changed. His examples do and mine did not rely on imagining what we would say were certain normal conditions of use to change.

Parallel examples and remarks can be developed about the other example Kripke relies on, that of gold. The key to constructing an example here is to refer to a variant metal that differs from gold not at all in its ordinary but only in certain hidden properties, thus blocking the appeal to our intuitions that is based on the large divergence in the practical properties of gold and iron pyrites. Suppose there is an initial fixing of reference of the term "gold" by a certain group of people. Their scientists later examine the original sample and discover that almost all items in it have the atomic number 79. The deviant part of the original sample is much closer to gold in its other properties than is fool's gold. The deviant items, and all other things of the same atomic structure, behave in every way of interest to the people exactly in the manner the non-deviant items and their like behave. The two types are for every practical purpose in the life of the people interchangeable, and only in the laboratory are the two distinguishable. Should we say with Kripke that this stuff without atomic number 79, but like the original non-deviant samples in every other way, is not gold?

I am inclined to say that it is gold and that there are two types of gold. But again the claim that it is not gold would make sense. It would mean in effect that it lacks a property that

most of the original samples had, namely that of being an element of atomic number 79. One could say either, depending on what criterion for "gold" one presupposed. But there is no justice in saying, as Kripke's theory requires, that the only right thing to say here is that the stuff in the deviant part of the sample is not gold. There are, however, some arguments in *Naming and Necessity* that seemingly attempt to prove exactly this point; I will examine them next.

IV

The following argument never surfaces with full explicitness but I think it may be one source of Kripke's strong feeling that his theory is correct. "The original concept of cat is: *that kind of thing*, where the kind can be identified by paradigmatic instances" (319). So too for gold, apparently. Now a certain hunk of metal is like gold in all practical respects, including all initial properties; but it does not have what almost all the paradigmatic instances had, namely atomic number 79. Because it so differs, this is *not the same kind of stuff* as that. If it were the same kind as that, then it would have the same nature; and it would therefore be of atomic number 79.

This appeal to the notion of "the same kind of stuff" or "the same nature" obviously gets us no further than the original appeal to what we would say about whether the stuff is gold or not. Contrary to the argument, it is possible to call this item the same kind of stuff as the paradigmatic instances. The items in question do fall within the extension of "the same kind of stuff". We can as readily call this the same kind of stuff as we can call this gold; and we have seen that it can be correct to call it gold. Everything depends on what we mean by, or count as, "the same kind of stuff". To insist that this relation holds only if the two samples in question have the same *k*-properties, for example same atomic number, is to presuppose a certain stipulation on the use of "the same kind of stuff". Kripke's theory can be read as resulting from just that stipulation. But if so, he cannot attempt to prove it correct by showing how it is supported by "intuitions" which presuppose that very stipulation.

Here is a related argument. "Even though we don't *know* the internal structure of tigers, we presuppose . . . that tigers form a certain species or natural kind. . . . We can say in advance that we use the term 'tiger' to designate a species, and that anything not of this species, even though it looks like a tiger, is not in fact a tiger" (§18).

Even if a people has the notion of a "natural kind", however, there is no necessity for them to circumscribe natural kinds along the same lines as Kripke's theory does. The two golds in the above example could reasonably be said to form a natural kind, even though they have different atomic numbers. Kripke's argument from the supposition that items form a natural kind suffers from the same weakness as the argument in terms of "being of the same kind". The natural use of "natural kind" doesn't carve up the conceptual territory in the way that Kripke wants it carved up; and if we want to use "natural kind" to carve it up that way we shall have to put a stipulation on its use. We will be stipulating that two things are of the same natural kind only if they have the same *k*-properties. Again, what seems to be an argument works only if one smuggles in a stipulative element that renders the conclusion true by fiat.

V

Kripke's arguments for his theory are inconclusive, and the theory itself is obscure and incomplete at some crucial points and subject to counterexamples at another. But is there a viable alternative to it? The criterial view of kind terms is such an alternative. In this section I shall show how it can deal with the scientific discovery of essence, by studying an example that at first appears to bear out Kripke's theory.

Simplifying history, we can say that viruses were first isolated as causal agents of certain diseases and as entities able to pass through barriers that normally filtered out microbes.

The reference of the term being fixed in this way, scientists tried to discover more about viruses. Professor Hughes, in *The Virus*, writes: "The problem [faced by scientists] was to discover the intrinsic properties of viruses rather than to characterize

them in terms of technique-determined ones [such as being filterable]".¹⁴

By intrinsic properties we can assume she means essential ones; thus here we seem to have an example that bears out Kripke's theory perfectly. A reference is fixed for the term "virus" and then scientists seek to discover the essence of these things.

She describes the denouement, as follows:

Electron microscopic and biochemical studies . . . indicated that a virus possesses an outer coat . . . of protein and an inner core of nucleic acid, either DNA or RNA, but not both. On the basis of this information viruses for the first time could be accurately defined and classified. . . . Modern definitions abound but most characterize the virus as an infectious, but not necessarily pathogenic, entity which is usually submicroscopic, which contains a core of either DNA or RNA covered by a protein or lipoprotein capsid and which reproduces exclusively within living cells.

I want to use this quotation to introduce a rival view of discovering essence, but not to make an argument from authority based on it. For first of all, the quotation can be read as supporting the realist view. It speaks of "accurately defining" the notion of a virus, and that can be read, of course, along Kripkean lines: The scientist discovers essence empirically, and an accurate definition of "virus" will be one that captures that discovered essence.

Nevertheless, a different account of discovering essence can be seen to fit Hughes's statement. That account focuses on the idea that scientists, in the case in question, and others, do put forward definitions. The rival view is this. Certain things called "*viruses*" are examined scientifically and found to have a number of properties, previously unknown. Some of the properties are utilized in constructing one or more definitions of what a virus is. At the point where a definition is given and accepted, essence is established. Essence is defined into being.

A crucial part of this view is that there are different possible definitions that might have been given, in the context in question. An important point made in the quotation is that there

14. Sally Smith Hughes, *The Virus* (Heinemann: London, 1971), p. 113.

are in fact a number of different definitions of "virus". When "virus" on a particular occasion of its use is used as bound by a given definition, then the essence of being a virus, on that use of "virus", is given by that definition.

We have now two rival pictures of how essence is discovered. Kripke's picture is this. A kind term is attached to an original sample when the reference of the term is first fixed. Scientists later examine either members of that sample or objects related in a certain way to the original sample. In doing so they discover certain scientific properties of the objects. (I think it is part of this picture that these are hidden properties.) In discovering these, essence is discovered.

The rival view introduces a step between the discovery of certain properties by science and the discovery of essence. Those discovered properties become essential ones by being incorporated by scientists in a definition of the natural kind. And other definitions are possible, given the discovered properties.

On the criterial view, the scientist discovers essence extensionally. This will happen when he discovers a certain property ϕ , which is then referred to in the definition of the kind in question. When and only when the definition is made does ϕ become an essential property. Such a definition states a criterial rule governing the use of "virus".

It is the element of definitional fiat involved, or sometimes involved, in discovering essence that Kripke's theory overlooks. There is nothing wrong in saying, in the example above, that the scientist has discovered the essence of the virus. But this discovery is not of the type that Kripke's theory envisaged. That is, one can call the extensional uncovering of essence a discovery of essence; but it is not what Kripke had in mind by discovering essence. That a discovered property is essential because defined or laid down to be so, or because incorporated into a definition or criterial rule, is something his theory does not allow. This definitional element goes against the spirit of Kripke's theory.

Hughes's use of the phrase "accurately defined" can be explained from the criterial viewpoint. With the extensional discoveries in question, accuracy of definition became possible in this sense: when the properties are known, definitions inconsis-

tent with the fact that viruses have those properties may be inaccurate. This still leaves open the question of whether a given discovered property is an essential one or not.

It is possible, consonant with the way the reference of "virus" was fixed, and consonant with the discoveries made by scientists about viruses, to define "virus" in some way or ways other than the way scientists actually did, once the properties in question were discovered. It is possible to discover that all known viruses, including those in the original sample, have an inner core of DNA or RNA and yet to refuse to incorporate that feature as a necessary condition for being a virus. One may wish, for theoretical reasons, or even as a hunch, to leave it open that entities be discovered in the future which are viruses but which lack this feature.

Here there is the following objection. If it hasn't been established scientifically that *all and only* viruses have this feature, then one hasn't really discovered essence. If on the other hand this generalization has been discovered, then one has no choice but to include the feature as part of the essence of being a virus.

But first, contrary to the objection, scientists could define "virus" in terms of a certain feature, and hence establish the essence of being a virus, even if no one had previously established the generalization that all and only viruses have this feature.

The second premise is trickier, and requires an examination of the idea of discovering that all and only viruses have this feature. We can discover that all and only the viruses *we know of* have this feature. We can also discover that the original set—the items picked out when "virus" was assigned a referent—have this feature. For the original set consisted, let us say, of the tobacco mosaic virus, and we know today that the tobacco mosaic virus has the feature. But how is one to go from such knowledge to the inference that "All and only viruses have this feature"? One would have to establish that if something is a virus it has this feature. But what is the criterion for being a virus? If the criterion is in terms of the definition employing the feature in question, the generalization will hardly count as a discovery. If it is in terms of some other definition, then that rival definition will delimit the conceptual bounds or essence of "virus". The

only alternative seems to be to rely on some mark of what a virus is, and upon an empirical generalization which states that whatever has that mark will also have the feature in question.

It is true that in this way we could establish as an empirical truth that all and only viruses have the feature. But it will still require a definition to raise the feature to the level of an essential feature. For, given the way that it was established that the feature holds of all viruses, it will be conceptually, or metaphysically, possible for us to discover instances of a virus which lacks the feature. It is logically possible to find something which bears the mark used in making the empirical generalization, but which lacks the feature. Kripke of course will deny this; so here is a point where the two systems come into conflict at a basic level. Assuming Wittgenstein's account of necessity the objection can be answered; assuming Kripke's account of necessity and essence, it cannot be answered.

To continue, from Wittgenstein's perspective if we decide to raise the feature of having a certain size and inner molecular structure, say, to the status of being an essential feature of viruses, then of course we will never in future discover a virus which lacks these features. But we may discover an entity in roughly the same size range, and of a similar molecular structure, and behavior; one that, for example, reproduces within cells and causes diseases similar in many respects to those known to be caused by viruses. Having adopted our terminology we will say, unless we decide to change, that these are not viruses; we will need a new term for them, whereas, if we had decided originally not to make the features in question essential ones, then we might well call these viruses. This illustrates the element of decision involved when a term is defined by means of properties science has discovered.

It will not do to object here: But these new things cannot be viruses because they are not the same kind of thing as the members of the original set. Whether they are the same kind, i.e. viruses, or not, depends on what terminology we have adopted. The fact that all the members of the original set had the features in question cannot force us to adopt these features as part of the definition of "virus". Nor is it true that our use of "the same kind" is tied, in such a case, to these features. The new

entities are of the same kind as those in the set, that is, they are viruses, provided we adopted one and not another of the many possible definitions of "virus" that we might have adopted. (Of course I am not saying that we are free to adopt just *any* definition of "virus".)

It might be argued that it is not in any way a matter of definition what, say, a virus is, because science itself will settle, in the long run, on what properly is to count as a virus; it will do so perhaps in terms of considerations about the simplicity and relative explanatory power of rival theories. This argument is invalid. It does not follow that, because science will itself somehow conclusively settle what a virus is, what a virus is is not a matter of definition. For it is possible that part of what science settles, in terms say of simplicity considerations, and so on, is what is the best definition of "virus".

VI

I will complete this account of the scientific discovery of essence by discussing two further aspects of Wittgenstein's notion of a criterion. These concern especially the pre-scientific use of kind terms. First, there is the idea that some terms are not governed by strict criterial rules. Second, the criteria governing a term sometimes change during the history of the use of the term. When such changes occur, the question arises of whether the meanings of the terms also change.

In the *Blue Book*, criteria are introduced in contrast to "symptoms".¹⁵ The difference is that where, say, it has been laid down by terminological fiat that disease D is the disease caused by virus V, to say "If the patient has the disease caused by V, then he has D" is a tautology, or necessary truth, whereas, given that we have noticed a correlation between the disease and a certain manifestation M, to say "If the patient manifests M, then he has disease D" is to make a hypothesis.

For Wittgenstein, statements of criteria are necessary because they are rules of language, whereas statements of symptoms are hypotheses made within a language governed in part by such

15. See the *Blue Book*, pp. 24-25 (*The Blue and Brown Books*, Harper: New York, 1958).

criterial rules. In the *Blue Book*, however, he goes on immediately to say that in natural language the distinction between criterion and symptom, or rule and hypothesis, is not always sharply drawn (p. 25). If a criterial rule for a term has been explicitly laid down by a speaker, a theorist, a committee on nomenclature, or whatever, then we can sharply distinguish, with respect to the term, criterial statements from noncriterial ones. Besides such explicitly made fiats, there may be cases where we can, by observation, distinguish clear-cut criterial rules from mere hypotheses. But in many cases there will have been no explicit fiat, and observation will not allow us to uncover criterial rules. A term is used; judgments using it are said to be true or false, and there is general agreement about such judgments. But no one can state a criterial rule governing such judgments, unless the person stating the rule makes an *ad hoc* decision to call a certain statement a (criterial) rule statement and others mere hypotheses. There will be different ways of making such a division. The language users may be persuaded to accept a criterial rule which someone fixes upon in an *ad hoc* way; but as Wittgenstein notes in the parallel discussion in the *Philosophical Investigations* (sec. 79) they may also easily be persuaded to accept a different one. The language as it exists does not dictate that one rather than another such division into symptoms and criteria is correct. In such cases we cannot sharply distinguish criteria from symptoms and at the same time think we are accurately stating how the people in question actually use language. Wittgenstein will say that they follow no strict rule.

When he refers to some criterion in the course of philosophizing, he is not to be read as always claiming that the term he is talking about is strictly governed by the criterion he alludes to, nor that it is strictly governed by any criterion. One use he makes of "criterion" is to draw precise lines where in nature there are only blurred ones. By his stating a criterion, we may be led to see more clearly how the associated term is used, even if the term is not in fact strictly governed by that criterion, in the sense that the criterion-symptom distinction is not sharply drawn by the people who use the term. An idealization or even a distortion of the use of a concept may be of value in getting us to understand how the concept is in fact used, in a way similar

to the manner in which a simplified, idealized, or even a distorted model of some cultural institution can help us understand it. (A caricature can draw our attention forcibly to certain features of the person caricatured.)

If one asks for the Wittgensteinian criterion governing "gold" as it was used, say, before the discovery of its atomic number, the response may be that the criterion-symptom distinction is not sharply drawn for this case. The people who use the term may not distinguish between criteria of gold and symptoms of gold. Such looseness, if it exists, is just a fact of their language; and any attempt to supply a strict criterial rule governing "gold" will then necessarily falsify its, in this respect, loose as opposed to strict use.

Usage can change. A term at one point bound by no strict criterial rule may have such a rule imposed upon it, as when the use of "grandmaster" in chess was formalized and strictly defined by F.I.D.E., the world chess federation. Also, a term bound by one strict criterion may have the criterion governing its use changed, as when F.I.D.E. changes one formal criterion for "grandmaster" for another. Sometimes the shift from no strict criterial rule to a strict one is a matter decided by science; and it can similarly adjudicate changes from one strict criterion to another. In the cases that we have been interested in, what seems to have occurred is a shift from a use of a term as bound by no strict criterion to one bound by a strict criterion introduced by science. Let us look at some examples of this.

We can suppose that prior to the discovery that purified water consists of H_2O , no one distinguished criteria for being water from "symptoms" for being water. Against this background one could introduce a criterion for being water, namely that it be H_2O . Typically in such a case both uses of "water" survive side by side. There is normally no confusion, because the context will indicate which use of "water" is in question. Thus when a sign on an exhibition at the Ontario Science Center says that the Toronto drinking water in the container is composed of a certain long and frightening array of substances, "water" is being used in the old way. On the other hand, uses of "water" governed by the criterion "Water is H_2O " are common in scientific contexts. On the present view, to say that it is of the es-

sence of water that it be H_2O , is to say that "water" (on the use in question) is governed by the criterion "Water is H_2O ". On this use to say that this is water because it is H_2O is to utter a definitional tautology.

"Virus" is also a term that changed from being not strictly governed by a criterion to being so governed. In the early stages of the investigation of viruses, it seems that there was no strict criterial definition of what a virus is. Some people spoke of viruses in certain circumstances, namely when they managed to isolate a disease-causing agent, which agent was not isolable as a cellular body. There was a controversy over whether there were indeed particle-like entities involved in these cases, or whether, for example, the disease was caused by a fluid-like substance having no particulate form. Various arguments showed, however, that the disease agent must be capable of reproduction in its hosts, and this was seen as ruling out the fluid hypothesis. The scientists here were to some degree arguing in the dark, in that they had no clear idea of what such particulate bodies or such fluids would be like if they existed. This managing without strict definitions is doubtless typical of science in certain stages of inquiry. "Virus" was governed by no strict criterion. With the isolation of particulate bodies by electron microscopy, it became possible to lay down an explicit criterial rule governing "virus", in terms of certain discovered properties, in much the way it is possible to lay down a criterion for being gold in terms of its atomic number.

An enormous sore point with this kind of account has been the question about change of meaning. Wittgenstein has been thought to be committed to the thesis that when there is a change of criterion (or equally, a change from no strict criterion to a strict one) then there must be a corresponding change of meaning. The idea of such a change of meaning has been vigorously opposed, especially by Putnam.¹⁶

But what is the point of contention? That is, what is at dispute when one speaks here of a change of meaning? It is true that Wittgenstein has spoken of change of meaning following

16. See, for example, "Dreaming and Depth Grammar," in R. J. Butler, ed., *Analytical Philosophy, First Series* (Blackwell: Oxford, 1966), 211-35.

upon a criterial change. But he did not mean to affirm what Putnam means to deny. As he uses "change of meaning" it is true *by definition* that criterial change entails meaning change. That tautology is presumably useful in prompting a certain way of looking at language. After all, something changes when the criterion changes. The expression is no longer used the way it was before. (Compare the two uses of "water" alluded to above—the one in the Science Center and the other in scientific literature.) Putnam says that in the sorts of cases Wittgenstein has in mind there really is no meaning change. The criterion governing "meaning change" in his assertion is not made explicit. It can only be the following. There is meaning change in a given case if, and only if, one's linguistic intuitions prompt one strongly to say in such a case that there is meaning change. There is nothing intrinsically wrong with adopting this criterion, but it is a mistake to think that, because by that criterion there is no meaning change in certain cases, therefore Wittgenstein is wrong about meaning change. His use is different.

Let us simply grant that on Putnam's use there is no meaning change in some, at least, of the cases that interest us. This removes one of the most strongly felt objections to Wittgenstein's view of criteria, without modifying that view in any essential way.

Given those observations what, more fully, does the discovery of essence look like from a criterial point of view? A kind term is in use in the language to refer to certain things, such as tigers, or a certain substance such as gold. It may be that there is no explicit distinction in the language between criteria governing the use of, say, "gold" and symptoms indicating that gold is present. We may be able to distinguish criteria from symptoms here only by means of an *ad hoc* decision. But there will be no problem in general in deciding that something is gold; we will normally be able to establish this with certainty. The scientist will have no difficulty in picking out instances of gold to serve as targets of scientific study. Scientists may discover, say, that gold is an element and has the atomic number 79. It is then possible to define gold as the element with that number. When that definition is adopted, then it is of the essence of gold (on the use of "gold" so governed) to be of that

atomic number; but whether to adopt that definition was a matter for decision, not discovery. The definition may have been only implicitly and not explicitly adopted; in that case, that it is the criterion governing "gold" will show itself in features of the corresponding use of "gold". Two uses of "gold" may continue to exist side by side, with the context of use serving to distinguish them. Parallel to the case of "water", there will be a common use of "gold" to refer to a metal that meets certain standards of purity, and a more restrictive use to refer to the pure element itself. We do not have to say that the meaning of "gold" changes when its criterion is fixed in the above way. Whether we say that or not will depend on what we count as establishing a "change of meaning". We can nevertheless fruitfully note that *something* has changed; a new rule has been introduced.

I have tried to say what Wittgenstein's view of kind terms really is, and to defend it against Kripke's criticisms. I view this paper as vindicating Malcolm's basic conception of a criterion, and as helping to justify something that underlies his work in philosophy—his early, long continued, and powerfully defended assumption of the importance of Wittgenstein.

The Perception of Shape

DAVID H. SANFORD

In this essay I shall expand on a response of Thomas Reid to an argument of Berkeley for the heterogeneity of tangible and visual objects. I will agree with Reid both that shapes could be perceived by senses other than sight and touch and that shapes could be perceived by sight without any accompanying perception of hue or brightness. Then I will examine responses to the Molyneux problem by H. P. Grice and Judith Jarvis Thomson and argue, contrary to Grice's claim, that certain kinds of breakdowns between tangible and visible shape are conceivable. I hope that my claims that certain outlandish little stories are coherent, besides provoking interest in those who think there are philosophic grounds for denying their coherence, will also contribute to an understanding of the perception of shape. I shall supply some historical background before quoting the passage from Reid which primarily concerns me although I shall not attempt here to come to terms with the theory of perception which backs up the passages quoted from Berkeley or to say much about the several questions about the psychology of blindness and the psychology of visual depth perception raised by Molyneux's problem.

Sections 121-46 of Berkeley's *An Essay Towards a New Theory of Vision* are devoted to supporting the proposition that

The extension, figures, and motions perceived by sight are specifically distinct from the ideas of touch called by the same names,

nor is there any such thing as one idea or kind of idea common to both senses. (NTV, section 127, Berkeley's italics.)

After presenting three arguments for this contention in sections 129-31, Berkeley writes (NTV, section 132):

A farther confirmation of our tenet may be drawn from the solution of Mr. Molyneux's problem, published by Mr. Locke in his *Essay*: Which I shall set down as it there lies, together with Mr. Locke's opinion of it, "Suppose a man born blind, and now adult, and taught by his touch to distinguish between a cube and a sphere of the same metal, and nighly of the same bigness, so as to tell, when he felt one and t'other, which is the cube and which the sphere. Suppose then the cube and sphere placed on a table, and the blind man to be made to see: *Quaere*, Whether by his sight, before he touched them, he could now distinguish and tell which is the globe, which is the cube?' To which the acute and judicious proposer answers: 'Not. For though he has obtained the experience of how a globe, how a cube, affects his touch, yet he has not yet attained the experience that what affects his touch so or so must affect his sight so or so: Or that a protuberant angle in the cube that pressed his hand unequally shall appear to his eye as it doth in the cube.' I agree with this thinking gentleman, whom I am proud to call my friend, in his answer to this his problem; and am of opinion that the blind man at first sight would not be able with certainty to say which was the globe which the cube, whilst he only saw them." (*Essay on Humane Understanding*, book II, chapter 9, section 8.)

The passage Berkeley quotes was added by Locke to the second edition of the *Essay* after receiving a letter of March 2, 1692/3, from William Molyneux of Dublin. Molyneux posed his problem earlier in a letter of July 7, 1688, which besides asking of the formerly blind man "Whether he could, by his sight, and before he touch them, know which is the Globe and which the Cube?" also asks "Whether he could know by his sight, before he stretched out his Hand, whether he could not Reach them, tho they were Removed 20 or 1000 feet from Him?"¹ Molyneux's

1. Although this letter, sent to the authors of the *Bibliothèque Universelle* and addressed to "the Author of the *Essai Philosophique concernant L'Entendement*," was found among Locke's papers, neither Locke nor Molyneux refers to it in their subsequent correspondence. The letter is discussed by

only condition is that the man born blind has never made a direct correlation between sight and touch before being asked to distinguish sphere from cube. He does not specify in either letter that the blind man be questioned on the very first occasion he can see, so Locke's addition "at first sight" significantly changes the problem.² It is consistent with Molyneux's formulation that the man born blind have any amount of time to become accustomed to seeing before being asked the question so long as he has never touched cubes or spheres while seeing them. Thus understood, Molyneux's problem suggests a variety of experiments which, so far as I know, have never been attempted. A man born blind who could be made to see and who had himself sufficient theoretical interest in the psychology of blindness and shape perception might be willing to refrain,

Désirée Park in "Locke and Berkeley on the Molyneux Problem," *Journal of the History of Ideas*, 30 (1969): 253-60. The letter makes clear that Molyneux thought of his problem as concerning the visual perception of depth or distance, a topic he wrote about elsewhere. Berkeley's argument that "Distance, of itself and immediately, cannot be seen" (*NTV*, section 2) is taken from Molyneux's *Dioptrica Nova*. Still, for Berkeley's purpose of denying the existence of common sensibles, Molyneux's question can be restricted to the identification of two-dimensional shapes. This is just what Berkeley does immediately after quoting Locke's version of the problem.

Now, if a square surface perceived by touch be of the same sort with a square surface perceived by sight, it is certain the blind man here mentioned might know a square surface as soon as he saw it. (*NTV*, section 133)

J. L. Mackie discusses the restriction of Molyneux's problem to two-dimensional shapes in his *Problems from Locke* (Oxford: Oxford University Press, 1976), pp. 30-31. A similar restriction was suggested by W. H. S. Monck in *Space and Vision* (1872). References to Monck's and many other responses to Molyneux's problem may be found in a useful brief historical survey by John W. Davis, "The Molyneux Problem," *Journal of the History of Ideas*, 21 (1960): pp. 392-408.

2. Molyneux seems never to have remarked on this change or on Locke's slight misquotation of his March 2, 1692/3, letter. Where Molyneux writes "cube and a sphere (suppose) of ivory," Locke has "cube and sphere of the same metal." If this means merely "of the same material," the change is insignificant. If we think of objects of metal rather than of ivory, however, the change increases the difficulty of the task set the man born blind. Since metal is more reflective than ivory, its polished surface acts more as a mirror, and mirror surfaces can be confusing even to persons of normal vision.

while he learns to see, from making any correlations between shapes he sees and shapes he can determine without seeing. Since he knows what shape a hand is, he should not be allowed even to hold his hands before his face while the bandages are off his eyes. If he can eventually learn names of the shapes he sees under such strictly controlled conditions, he is taught names unknown to him before he was made to see and never applied to objects he feels while his eyes are rebandaged. Should he in this way come easily to identify a *Kugel* and a *Würfel* by sight, perhaps only after many hours of visual training, he can then be asked whether a *Kugel* is a cube and a *Würfel* a sphere, or the other way round.

In his *An Inquiry into the Human Mind*, Reid holds that "sight discovers almost nothing which the blind may not comprehend" (title of section 2, chapter 6, "Of Seeing"). Although he does not mention the Molyneux problem by name, he is clearly concerned to reject Berkeley's answer to it. Reid writes of the famous blind geometer at Cambridge that

. . . if Dr. Saunderson had been made to see, and attentively had viewed the figures of the first book of Euclid, he might, by thought and consideration, without touching them, have found out that they were the very figures he was before so well acquainted with by touch. (Section 11.)

Reid is also concerned to reject Berkeley's arguments. An argument in *NTV*, section 127, proceeds from the assumption that "light and colours are allowed by all to constitute a sort of species entirely different from the ideas of touch." Since "there is no other immediate object of sight besides light and colours," it is "a direct consequence, that there is no idea common to both senses." Against the suggestion that figure and extension are immediate objects of sight in addition to light and color, Berkeley writes,

And as for figure and extension, I leave it to anyone that shall calmly attend to his own clear and distinct ideas to decide whether he has any idea intromitted immediately and properly by sight save only light and colours: Or whether it be possible for him to frame in his mind a distinct abstract idea of visible extension or figure exclusive of all colour; and on the other hand,

whether he can conceive colour without visible extension? For my own part, I must confess I am not able to attain so great a nicety of abstraction. . . .³

Although he does not use the philosophical term 'abstract idea', Reid attempts to answer all these challenges of Berkeley in the following three paragraphs, which I number for future reference.⁴

[1] Let us suppose, that the eye were so constituted, that the rays coming from any one point of the object were not, as they are in our eyes, collected in one point of the *retina*, but diffused over the whole: it is evident to those who understand the structure of the eye, that such an eye as we have supposed, would shew the colour of a body as our eyes do, but that it would neither shew figure nor position. The operation of such an eye would be precisely similar to that of hearing and smell; it would give no perception of figure or extension, but merely of colour. Nor is the supposition we have made altogether imaginary: for it is nearly the case of most people who have cataracts, whose crystalline, as Mr. Cheseldon observes, does not altogether exclude the rays of light, but diffuses them over the *retina*, so that such persons see things as one does through a glass of broken jelly; they perceive the colour, but nothing of the figure or magnitude of objects.

[2] Again, if we should suppose, that smell and sound were conveyed in right lines from the objects, and that every sensation of hearing and smell suggested the precise direction or position of its object; in this case the operations of hearing and smelling would be similar to that of seeing: we should smell and hear the figure of objects, in the same sense as now we see it; and every smell and sound would be associated with some figure in the imagination, as colour is in our present state.

3. NTF, section 130. This passage is discussed by George Pitcher in *Berkeley* (London: Routledge & Kegan Paul, 1977), pp. 53-55.

4. I doubt if it is necessary to respond to these challenges in the extreme way Reid does to resist Berkeley's argument for the heterogeneity of tangible and visible figure, a purpose better served by pointing to the distinction illustrated by an ambiguous sentence such as "Visible figure cannot exist apart from light and color." This can mean either (1) if something is a figure, it cannot be visible apart from light and color, or (2) if something is a visible figure, it cannot exist apart from light and color. The fact that (1) does not imply (2) damages Berkeley's argument, and of course this fact cannot be appreciated until (1) and (2) are distinguished.

[3] We have reason to believe, that the rays of light make some impression upon the *retina*; but we are not conscious of this impression; nor have anatomists or philosophers been able to discover the nature and effects of it; whether it produces a vibration in the nerve, or the motion of some subtile fluid contained in the nerve, or something different from either, to which we cannot give a name. Whatever it is, we shall call it the *material impression*; remembering carefully, that it is not an impression upon the mind, but upon the body; and that it is no sensation, nor can resemble sensation, any more than figure or motion can resemble thought. Now, this material impression, made upon a particular point of the *retina*, by the laws of our constitution, suggests two things to the mind, namely, the colour, and the position of some external object. No man can give a reason, why the same material impression might not have suggested sound, or smell, or either of these, along with the position of the object. That it should suggest colour and position, and nothing else, we can resolve only in our constitution, or the will of our Maker. And since there is no necessary connection between these two things suggested by this material impression, it might, if it had so pleased our Creator, have suggested one of them without the other. Let us suppose, therefore, since it plainly appears to be possible, that our eyes had been so framed, as to suggest to us the position of the object, without suggesting colour, or any other quality: what is the consequence of this supposition? It is evidently this, that the person endued with such an eye, would perceive the visible figure of bodies, without having any sensation or impression made upon his mind. The figure he perceives is altogether external; and therefore cannot be called an impression upon the mind, without the grossest abuse of language. If it should be said, that it is impossible to perceive a figure, unless there be some impression of it upon the mind; I beg leave not to admit the impossibility of this, without some proof: and I can find none. . . .³

5. *Inquiry*, VI, section 8. This passage appears on pp. 117-19 of Timothy J. Duggan's edition of Reid's *Inquiry* (Chicago: University of Chicago Press, 1970) and on pp. 145-46 of Hamilton's edition. Hamilton discusses it critically but not, I think, very usefully, in Note E. "On the Correlative Apprehensions of Colour, and of Extension and Figure," pp. 917-20. The passage is discussed briefly by Norman Daniels, *Thomas Reid's Inquiry: The Geometry of Fisiiles and the Case for Realism* (New York: Burt Franklin, 1974), pp. 84-86. Daniels says he is not concerned "with any effort either to support or to attack Reid on the merits of his thought experiment" but rather with Reid's reduction "to a minimum the role played by sensations in concept formation" (p. 86).

Reid's suggestion in paragraph [1] that one can perceive color without any shape being differentiated within one's visual field is supported by the common enough experience of seeing a uniformly colored expanse which is large enough or close enough to fill one's visual field completely. But Reid suggests something stronger, that one can perceive color without perceiving extension. Examples of diffusion of light, by cataracts, glasses of broken jelly, or anything else, are insufficient to support this suggestion since, in considering these examples, we still think of the subject's having an extended visual field. It is as if, we think, the subject were seeing an undifferentiated colored expanse, which is still an expanse, and is thus extended, no matter how indeterminate its boundaries are.⁶ I believe that a different sort of example will give better support to Reid's stronger suggestion. This example can be discussed more easily after considering paragraph [2].

In paragraph [2], Reid does not spell out how sensations of hearing and smell are to suggest precise direction or position when smell and sound are conveyed in right lines from the objects. I assume he is thinking of auditory and olfactory sense organs which are structurally analogous to the human eye. Since sounds emitted and reflected by objects do pretty much travel outwards in straight lines, while odoriferous particles diffuse aimlessly through the air, we will venture less far from the actual world if we consider only the auditory analogue of the eye. Such a sense organ would gather and focus sound waves on a surface analogous to the retina in having many receptors. If we do not imagine any great differences in the physical properties

In his later *Essays on the Intellectual Powers of Man*, Reid writes that "... space, whether tangible or visible, is not so properly an object of sense, as a necessary concomitant of the objects both of sight and touch" (*Essay II*, chapter 19, p. 324 in Hamilton's edition). I do not know how to show that the apparent conflict between this and the suggestion in paragraph [1] is merely apparent.

6. This point is made by John Immerwahr in his review of Daniels's book in the *Journal of the History of Philosophy*, 14 (1976): pp. 371-74, and by David A. Tebaldi in "Thomas Reid's Refutation of the Argument from Illusion," *Thomas Reid: Critical Interpretations*, edited by Stephen F. Barker and Tom L. Beauchamp (Philadelphia: Philosophical Monographs, 1976), p. 34, footnote 15.

of sound, then, because it has a wave-length much longer than light's, we have to imagine this multi-receptor ear to be much larger than a human ear if it is to register much information about the shapes of objects.⁷ (The distinction between imaginary human beings who have sense organs very different from ours and imaginary creatures whose sense organs are so different from ours that they should not be considered human is, I hope, unimportant for my enterprise. I assume that we could discuss perception with the creatures who have the strange kinds of sense organs I imagine.)

If imagining a multi-receptor organ of hearing supports Reid's suggestion in paragraph [2], then imagining a single-receptor organ of sight supports his suggestion in paragraph [1]. Instead of imagining light evenly diffused over the retina, with its thousands of light sensitive receptors, we can imagine eyes which have but a single receptor sensitive both to brightness and hue. The lens of such an eye could collect light and focus it on the receptor, but there would be no function served by its producing a differentiated retinal image. I think there is a reason for doubting that perceiving hue and brightness with such an eye would be like having a visual field constricted to a single point, a minimum visible. Suppose that a creature with multi-receptor ears is asked to imagine that hearing might convey information about pitch and loudness of sound without any accompanying perception of audible figure or extension. "Yes," the creature responds, "that would be like the constriction of the auditory field to a single point, a minimum audible." We protest. The creature with the multi-receptor ear is being asked to imagine an ear just like ours, a single-receptor ear. Our auditory experience is not of an otherwise empty auditory field constricted to a point. We need not suppose, therefore, that the possessor of a single-receptor eye would perceive points of light or have a punctiform visual field. We may well doubt that we know what it would be like to perceive hue and brightness without perceiving anything as a colored expanse or as a point of light, but so might the owners of multi-receptor ears doubt

7. Without reference to Reid, W. C. Clement discusses the possibility of hearing surfaces in "Seeing and Hearing," *The British Journal for the Philosophy of Science*, 6 (1955): pp. 61-65.

that they could imagine what it would be like to hear with ears like ours.

When Reid attempts to imagine hearing and smelling shapes, he considers how this might be analogous to seeing shapes. I will presently consider how smelling shapes might be analogous to perceiving shapes by touch. This exercise should help us understand our actual perception of shape by the sense of touch, for we will have to decide what the relevant features of this process are before we can decide how perception by another sense could be analogous to it. Consideration of the perception of shape by the sense of touch will also reveal how the sensations of which we are conscious in perceiving the shape of something are often insufficient for a judgment of shape to be based on them or to be inferred from their felt characteristics, and this point will bear on Reid's contention in paragraph [3] that we could perceive shapes visually without any sensations of brightness or hue.

If you hold a basketball between your hands with your fingers, thumbs, and palms all touching the ball, you normally can feel, without looking, that (1) your hands are making maximum contact with portions of the surface of an object. You can also feel both that (2) *the portions of the surface of the object you are holding are convex and are a certain distance apart, about the diameter of a basketball*, and that (3) your hands are curved toward each other and are separated by that distance. It follows from (1) that (2) is true if and only if (3) is true. If you should ask yourself how you know that (3) is true, it may seem natural to appeal to (1) and (2); but if you should ask yourself how you know that (2) is true and you make no use of information you *have about the shape and size of the object you are holding* that you cannot determine from your present experience of holding it, you will appeal to (1) and (3). Together the two accounts appear to be involved in a tight circle: "I know my hands are in a certain position because I know they are in contact with an object of a certain shape and size, and I know the object is of this shape and size because I know my hands are in a certain position and in contact with it." The appearance of circularity can be diminished by distinguishing ways of understanding a ques-

tion of the form "How do you know that p ?" If it is understood as a request for information of which you are conscious and from which you can validly infer that p , then either answer can be appropriate and the first can seem more natural because the fact that you are touching an object of a certain size and shape is normally of more interest, and is thus more likely to be consciously considered, than the fact that your hands are in a certain position. If the question is understood as asking about what information you are receiving and processing, whether consciously or not, in order to possess the information that p , then the two answers do involve a genuine circularity, and the first answer has things backwards.⁸

Now consider how you would move your hands over the surface of the basketball to determine its overall shape. As before, the fact that the object you are holding is a sphere is more likely to be consciously considered than the fact that your hands trace a spherical shape as they move in contact with the ball. Still, information about the relative positions of your hands, whether consciously considered or not, is processed somehow in your perception of the shape of the object you hold. Perception of contact between a part of your body and something else, on the other hand, is not normally necessary to determine the relative positions of our limbs and shapes of the spatial paths they trace as we move them. The shapes of these spatial paths play an important role in our perception of boundaries by touch.

In chapter 5 of the *Inquiry*, "Of Touch," Reid writes,

We are commonly told by philosophers, that we get the idea of extension by feeling along the extremities of a body, as if there was no manner of difficulty in the matter. I have sought, with

8. Reid makes the same sort of point.

When my two hands touch the extremities of a body; if I know them to be a foot asunder, I easily collect that the body is a foot long; and if I know them to be five feet asunder, that it is five feet long; but if I know not what the distance of my hands is, I cannot know the length of the object they grasp; and if I have no previous notion of hands at all, or of the distance between them, I can never get that notion by their being touched. (*Inquiry*, V, section 6, pp. 74-75 of the Duggan edition.)

great pains I confess, to find out how this idea can be got by feeling, but I have sought in vain. (Section 5, p. 71 in Duggan's edition.)

Part of the explanation of Reid's failure must be that we seek in vain when we attempt to attend to the bodily sensations upon which we might suppose our knowledge of the positions and movements of limbs must be based. Sensory receptors in our joints are largely responsible for our acquiring this knowledge, but they usually do not supply us with an awareness of sensations which are sufficiently differentiated to account for the precision and accuracy of the judgments we can make about positions and motions of various parts of the body. G. E. M. Anscombe applies the phrase 'non-observational knowledge' to this phenomenon.⁹ Wittgenstein discusses the phenomenon as follows:

"My kinaesthetic sensations advise me of the movement and position of my limbs."

I let my index finger make an easy pendulum movement of small amplitude. I either hardly feel it, or don't feel it at all. Perhaps a little in the tip of the finger, as a slight tension. (Not at all in the joint.) And this sensation advises me of the movement?—for I can describe the movement exactly.¹⁰

When the tip of the finger traces the edge of a coin, it is no wonder that we are unaware of sensations which account for our perception of the coin's shape, since we are unaware of sensations which account for our ability to describe the circular movement of the finger tip. Our ability to perceive the coin's shape by touch depends on our ability somehow to receive and process information about the movement of the finger. In the language of Reid's paragraph [3], the sense receptors in the joints of the finger receive a *material impression* from movement of the joint. That these material impressions enable us to describe precisely the movements of the finger without producing correspondingly precisely differentiated sensations "we can resolve only in our

9. *Intention* (Oxford: Basil Blackwell, 1958), section 8, pp. 13-15.

10. *Philosophical Investigations* (Oxford: Basil Blackwell, 1953), part II, section viii, p. 183c. Notes on a lecture by Wittgenstein on this topic are quoted by Norman Malcolm in *Ludwig Wittgenstein: A Memoir* (London: Oxford University Press, 1958), pp. 48-49.

constitution." It would be in the spirit of paragraph [3] to go on to imagine perception of shape by touch in which the subject has even less awareness of what goes on in the process of perception. If information about contact between the finger and the coin and information about the movement of the finger is received and processed and results in the judgment that one is touching a circular disc, why, Reid could ask, need the perceiver be able to describe the motion of his finger, or why need he be aware of any sensation in the finger tip which indicates contact between the finger and the coin's edge?

The shape of a thing with boundaries is determined by its boundaries, and its boundaries are determined by a more or less discontinuous change through space with respect to some property. Changes in the relational property of impenetrability mark the boundaries of the objects whose shapes we perceive by touch. Changes in the relational properties of light emission or light reflection mark the boundaries of objects we perceive by sight. Not everything with a shape has a boundary in this manner. The orbits of the planets, for example, are elliptical, but the shape of an orbit is determined by the path of the planet through space rather than by a discontinuity of some spatial property. We have seen how this kind of shape, the shape of a path of something (such as our hands) moving through space, is important to our perception of shape by the sense of touch.

Discontinuities in spatial properties may be imperceptible for many sorts of reasons. Some of these discontinuities can be detected only indirectly via the perception of something else, as when the directly detected boundary marked by a contrast between darker and less dark is taken to indicate the coincident boundary of the portion of a necktie affected by a splatter of grease. Among the things upon which the direct detection of discontinuities depends is the absence of obstacles between the region of discontinuity and the sense organ. Some fairly definite boundaries marked by differences in directly detectable sensible qualities are usually obscured by other things. Often such an obstacle determines the shape of the boundary it hides, as when the interior shape of an opaque bottle determines the shape of the liquid-nonliquid boundary that the walls of the bottle conceal from view.

Consider a sound-proof and air-tight room filled with sound and odor. The space enclosed by the room is qualified by auditory and olfactory qualities. The walls of the room, which interfere with the direct detection of the boundaries marked by discontinuities of olfactory and auditory qualities, also determine the shape of these boundaries. Although the shape of the space filled with sound and odor is thus causally determined by the shape of the room, it is logically coherent to suppose that the auditory and olfactory boundaries should be independent of a container. If we imagine that the walls of the room disappear without any change in the qualification of spatial regions by auditory or olfactory qualities, we thereby imagine more or less definite auditory or olfactory spatial boundaries which are not causally dependent on the boundaries of more familiar physical objects. With boundaries marked by impenetrability out of the way, the olfactory and auditory boundaries become approachable. Without attempting to imagine possible physical explanations for these boundaries being fairly stable without the stability of ordinary physical objects to determine them, let us imagine how these boundaries might be perceived in a way analogous to our perception by touch.

Our perception of boundaries by touch depends on these being the same kind as the boundaries of our bodies. Material bodies are mutually impenetrable, and it is the more or less discontinuous increase in difficulty in moving a part of the body in a certain direction which conveys the information that our hand, glove, screwdriver, barge pole, or whatever, has come into contact with another body. The auditory and olfactory boundaries we are imagining can be crossed without difficulty by whatever new sense organs we will imagine, such as a long, highly mobile olfactory organ (a "feeler") which reaches out in search of spatial discontinuities of olfactory qualification. A sudden increase or decrease in some olfactory quality would indicate that the sensitive part of the feeler has crossed a boundary. A single-receptor feeler would have continually to cross and recross the boundary to convey the information that the receptor is close to the boundary as the whole feeler moves in a larger pattern to trace the overall shape of the bounded region. A feeler with several receptors a short distance apart could stay in the immediate

vicinity of the boundary without so much back and forth motion since the difference in quality which marks the boundary could be detected by one receptor being affected differently from its near neighbor. And we can also imagine creatures with a number of such feelers which explore the environment simultaneously.

This exercise in imagining how shapes could be perceived by a sense different from touch and sight is not a phenomenological one. I am attempting to describe how information sufficient to form correct judgments about shape might be received by other senses, not what it would be like to perceive shape by these senses. It is obvious that information about the position of the feelers I have described would play some role in forming judgments about the shapes and sizes of the boundaries these feelers detect, just as information about the positions of parts of our bodies plays a role in our perception of shape by the sense of touch. In the spirit of Reid's paragraph [3], we can ask questions about the operation of these imagined senses similar to those we asked about perception by touch. Need creatures which use these strange senses be aware of sensations in their feelers which enable them to judge the positions and movements of their feelers? Need they be able to judge the positions and movements of their feelers to be able to judge the shapes of boundaries perceived by the feelers? Need they be aware even of any sensations which indicate the presence or absence of a sensible quality to detect boundaries marked by differences in such a quality?

Reid's suggestion in paragraph [3] is the most radical of the lot. Those who have discussed it have mostly thought it could not be correct. Dugald Stewart, keeping pretty close to Reid's own words in a slightly later passage, puts it as follows:

Our eye *might* have been so framed as to suggest the figure of the object, without suggesting colour, or any other quality; and, of consequence, that there seems to be *no sensation* appropriated to visible figure; this quality being suggested *immediately* by the material impression upon the organ, of which impression we are not conscious.

To this, Stewart replies,

Consider a sound-proof and air-tight room filled with sound and odor. The space enclosed by the room is qualified by auditory and olfactory qualities. The walls of the room, which interfere with the direct detection of the boundaries marked by discontinuities of olfactory and auditory qualities, also determine the shape of these boundaries. Although the shape of the space filled with sound and odor is thus causally determined by the shape of the room, it is logically coherent to suppose that the auditory and olfactory boundaries should be independent of a container. If we imagine that the walls of the room disappear without any change in the qualification of spatial regions by auditory or olfactory qualities, we thereby imagine more or less definite auditory or olfactory spatial boundaries which are not causally dependent on the boundaries of more familiar physical objects. With boundaries marked by impenetrability out of the way, the olfactory and auditory boundaries become approachable. Without attempting to imagine possible physical explanations for these boundaries being fairly stable without the stability of ordinary physical objects to determine them, let us imagine how these boundaries might be perceived in a way analogous to our perception by touch.

Our perception of boundaries by touch depends on these being the same kind as the boundaries of our bodies. Material bodies are mutually impenetrable, and it is the more or less discontinuous increase in difficulty in moving a part of the body in a certain direction which conveys the information that our hand, glove, screwdriver, barge pole, or whatever, has come into contact with another body. The auditory and olfactory boundaries we are imagining can be crossed without difficulty by whatever new sense organs we will imagine, such as a long, highly mobile olfactory organ (a "feeler") which reaches out in search of spatial discontinuities of olfactory qualification. A sudden increase or decrease in some olfactory quality would indicate that the sensitive part of the feeler has crossed a boundary. A single-receptor feeler would have continually to cross and recross the boundary to convey the information that the receptor is close to the boundary as the whole feeler moves in a larger pattern to trace the overall shape of the bounded region. A feeler with several receptors a short distance apart could stay in the immediate

vicinity of the boundary without so much back and forth motion since the difference in quality which marks the boundary could be detected by one receptor being affected differently from its near neighbor. And we can also imagine creatures with a number of such feelers which explore the environment simultaneously.

This exercise in imagining how shapes could be perceived by a sense different from touch and sight is not a phenomenological one. I am attempting to describe how information sufficient to form correct judgments about shape might be received by other senses, not what it would be like to perceive shape by these senses. It is obvious that information about the position of the feelers I have described would play some role in forming judgments about the shapes and sizes of the boundaries these feelers detect, just as information about the positions of parts of our bodies plays a role in our perception of shape by the sense of touch. In the spirit of Reid's paragraph [3], we can ask questions about the operation of these imagined senses similar to those we asked about perception by touch. Need creatures which use these strange senses be aware of sensations in their feelers which enable them to judge the positions and movements of their feelers? Need they be able to judge the positions and movements of their feelers to be able to judge the shapes of boundaries perceived by the feelers? Need they be aware even of any sensations which indicate the presence or absence of a sensible quality to detect boundaries marked by differences in such a quality?

Reid's suggestion in paragraph [3] is the most radical of the lot. Those who have discussed it have mostly thought it could not be correct. Dugald Stewart, keeping pretty close to Reid's own words in a slightly later passage, puts it as follows:

Our eye *might* have been so framed as to suggest the figure of the object, without suggesting colour, or any other quality; and, of consequence, that there seems to be *no sensation* appropriated to visible figure; this quality being suggested *immediately* by the material impression upon the organ, of which impression we are not conscious.

To this, Stewart replies,

To my comprehension, nothing can appear more manifest than this, that, if there had been no variety in our sensations of colour, and still more, if we had had no sensation of colour whatsoever, the organ of sight could have given us no information, either with respect to *figures* or to *distances*; and, of consequence, would have been as useless to us, as if we had been afflicted, from the moment of our birth, with a *gutta serena*. (*Collected Works*, Hamilton edition, I, pp. 132-33, footnote.)

Stewart's response is a natural one because of the extreme difficulty of imagining what it would be like to see with eyes of the sort Reid describes. Yet his principle that the organ of sight cannot convey information without producing visual sensations should appear to be questionable when it is recognized as an instance of the general principle that a sense organ cannot convey information without producing characteristic sensations. Our semicircular canals certainly convey information, although there are no characteristic sensations produced by the vestibular system. The sense receptors in our joints certainly convey information, although the sensations they produce do not play an important role in our perception of position and movement.

Reid's thought experiment in paragraph [3] does not involve imagining any changes in the production of the retinal image, or material impression. Some variety in visual qualities which mark the boundaries of the seen shapes is still necessary for the production of the corresponding retinal image. Reid supposes that there could be perception of the boundaries marked by differences in visual quality without any perception of the qualities themselves. Study of recent experimental work in sensory substitution should diminish the inclination to think that this is impossible.

Paul Bach-y-Rita and Carter C. Collins have experimented with a "Tactile Visual Substitution System" (TVSS).

The essential idea behind TVSS is to throw an image of objects on to the skin of the blind person, using electrically driven vibrators to provide the stimulation. An array of some four hundred of these vibrators is attached to a ten-inch square of skin on the back or abdomen of the perceiver. A television camera, which the person can point where he desires, sends electrical signals to the vibrators, causing some of them to vibrate, depending on the

object in front of the camera. Each vibrator covers a small area of the image captured by the camera, much as a newspaper photograph represents a scene by an array of dots.¹¹

Subjects experience sensations on the skin when the camera is motionless, when the vibrators produce discomfort, and when they first start learning to use the system. Subjects who have learned to move the camera report experiencing shapes in front of them.

With training, the blind subjects can identify and correctly locate in space complex forms, objects, figures, and faces. Perspective, parallax, size constancy, including looming and zooming and depth cues, are correctly utilized. The subjective localization of the information obtained through the television camera is not on the skin; it is accurately located in the three-dimensional space in front of the camera, whether the skin stimulation matrix is placed on the back, on the abdomen, on the thigh, or changed from one of these body locations to another.¹²

Although information is conveyed via tactile receptors, received by routes independent of the optic nerve, and not processed in the visual cortex, it is not the quality of the tactile sensations that accounts for the subject's ability to perceive shape. The information carried by the light passing through the lens of the television camera is of just the same sort as the information received by a normal eye. It can thus reasonably be called *visual* information and the TVSS can be regarded as providing a prosthetic eye.¹³

11. This summary of TVSS is taken from the final section of Michael J. Morgan's *Molyneux's Question* (Cambridge: Cambridge University Press, 1977), pp. 200-201. Earlier in that section he says, "When no definite solution is found to a question that has been debated for almost three hundred years, either it is meaningless or some new approach to it is wanting" (p. 198). Morgan suggests that TVSS may offer a new approach to the Molyneux problem.

12. Paul Bach-y-Rita, *Brain Mechanisms in Sensory Substitution* (New York: Academic Press, 1972), p. ix.

13. Not every means by which a blind subject can acquire information about the shapes and locations of distant objects thereby involves visual information. Blind persons who find their way around by detecting sound echoes do so by processing a different sort of information. In order to diminish further the natural supposition that subjects who use a TVSS must be consciously noting

"Do the blind really see with a TVSS?" asks Morgan. His emphatic answer will not convince the sceptic.

Either they do, or psychology as a science is impossible. After all, they are being given the same stimulation that causes the sighted to see, and they are giving the same responses. If despite the identity between both input and output they are having different 'perceptions', and if we can never know whether they are or not, then perception is not a fitting subject for a scientific work. In fact it is, as we all know, so I conclude that the TVSS allows the blind to see (and not merely to 'see'). (Morgan, p. 204.)

The subject using a TVSS may give the same responses as the normal viewer to certain limited questions about the shapes and locations of objects in front of him, but so might a blind man equipped with a device that whispered the correct answer in his ear. If we count sincere reports of what the perceiver's experience is like as a response or "output", the output of a subject using a TVSS is scarcely identical to that of a person with normal vision. Still, there are impressive similarities to normal seeing unaccompanied by additional similarities to familiar non-visual ways of acquiring information. I do not know if a blindfolded person with normal vision has learned to use a TVSS. Although it is natural to think that if we acquired the same competence with a TVSS as the most accomplished blind subjects,

a character of their tactile experience, we should note that subjects who use their sense of hearing to navigate do not do so by consciously noting a character of their auditory experience. Indeed, they are often unaware even that the sense of hearing is involved.

When questioned as to how they do it, they reply that as they approach a large obstacle such as a wall, they begin to feel a very light pressure against the skin of their faces, and that this pressure increases as they get nearer the wall.

We now know that the stimuli for these sensations are not pressure on the skin or any other kind of cutaneous stimuli, but auditory stimuli in the shape of the echoes of the subject's own footsteps, or the tapping of his stick on the pavement reflected back to him by the obstacle. The question naturally arises: how does this auditory stimulus fail to be perceived as auditory and to be perceived as cutaneous? (J. Taylor, *The Behavioral Basis of Perception* (New Haven: Yale University Press, 1962), quoted by Morgan, p. 167.)

we would then be able to say whether or not we and they really see with its help, such an experiment could be inconclusive. The experience with a TVSS is surely different from the experience of normal seeing, if only because the amount of information transmitted by the television camera is far less than the amount of information transmitted by a normal retina. We may think that since it is in virtue of differences in brightness that information is received by the television camera, the subject must experience something like differences in brightness and must thereby experience something like a figure-ground phenomenon, with shapes standing out from a darker or lighter background. One subject who has learned to perceive shapes with TVSS flatly denies that his experience is anything like this.¹⁴ Reid did not anticipate anything like the details of TVSS, but I think he did "beg leave not to admit the impossibility" of what TVSS shows to be possible. Someone can perceive a white shape against a dark background without being able to perceive white or dark. Boundaries determined by a change in spatial quality can be perceived without perceiving the qualities which determine the boundaries. Reid's own case is definitely a case of seeing. It involves no sensory substitution, and it is compatible with the perceiver's processing all the visual information, while not being able to make all the same visual judgments, as a person with normal vision.

I postpone the drawing of further conclusions from the discussion of Reid until I examine another sort of response to the Molyneux problem.

At the beginning of her article "Molyneux's Problem,"¹⁵ Judith Thomson quotes Locke's version of Molyneux's 1692/3 letter, picks up Molyneux's phrases "affects his touch so or so" and "affect his sight so or so", and suggests that Molyneux is concerned with the possibility of divergent objects, objects whose tangible shapes differ from their visual shapes.

14. Personal communication with Gerard Guarniero, a blind subject who has published several articles on his experience with TVSS and has written a Ph.D. dissertation entitled "The Senses and the Perception of Space" (Department of Philosophy, New York University, 1977). Guarniero's first article is discussed by Morgan, pp. 205-7.

15. *The Journal of Philosophy*, 71 (1974) pp. 637-50.

I am inclined to think that what was in Molyneux's mind, because of which he drew that conclusion from those premisses, was this: that what affects one's touch so or so *could* have affected one's sight such and such instead of so or so—i.e., that what feels so or so could have looked such and such instead of so or so. Thus, for example, that what standardly feels like a globe could have standardly looked like a cube instead of like a globe, and vice versa. Certainly anyway, if you did think this, it would seem to you plausible to say that if a man hasn't had the experience of how things that feel so or so look, he can't tell by sight, i.e., from how a thing looks, how it feels. And plausible, then, to reason, as Molyneux does, that, if a man hasn't had the experience of how things that feel so or so look, he can't tell by sight, i.e., from how a thing looks, whether it feels like, and so is, a globe, or whether it feels like, and so is, a cube. The best he can do is guess. (pp. 637–38)

I doubt that Molyneux had the possibility of divergent objects in mind. If what standardly felt like a cube looked like a globe, it would not present the visual appearance of protuberant angles, since globes do not have protuberant angles. Molyneux, however, says that the man born blind and made to see "had not yet attained the experience . . . that a protuberant angle in the cube, that pressed his hand unequally, shall appear to his eyes as it does in the cube." He evidently assumes here that the cube does present the visual appearance of a protuberant angle to the man born blind although the man cannot identify it as the appearance of a protuberant angle.

The putative possibility Thomson intends to describe is very radical: given *any* object which can be both seen and felt, it standardly feels like a globe if and only if it standardly looks like a cube, and it standardly looks like a globe if and only if it standardly feels like a cube. As she points out, if shapes were thus universally divergent it is impossible that visually and tactually perceived spatial relations between shapes should be preserved. Consider the two-dimensional analogue. Four squares of equal size can be fit together to make a larger square. If these are all tangible squares which look to be circular, the visible circles cannot be seen to be related to each other as the tangible squares are felt to be related to each other since no four circles can be fit together to make a larger circle.

H. P. Grice has entertained the possibility of a much more limited breakdown between tangible and visual shape.¹⁶ He introduces the topic in a discussion of some Martians who in physical appearance

are more or less like ourselves, except that in their heads they have, one above the other, two pairs of organs, not perhaps exactly like one another, but each pair more or less like our eyes: each pair of organs is found to be sensitive to light waves. It turns out that for them x-ing is dependent on the operation of the upper organs, and y-ing on that of the lower organs. The question which it seems natural to ask is this: Are x-ing and y-ing both cases of seeing, the difference between them being that x-ing is seeing with the upper organs, and y-ing is seeing with the lower organs? Or alternatively, do one or both of these accomplishments constitute the exercise of a new sense, other than that of sight? (p. 146)

The Martians learn and use our color vocabulary, but they insist that there is a big difference between x-ing blue and y-ing blue. Grice uses the story of the Martians to support his claim that some reference to the special introspectable characteristic of the experience of perceiving by a certain sense is necessary to distinguish perceiving by that sense from perceiving by another sense.¹⁷ He considers the suggestion that even if x-ing and y-ing are exercises of different senses because of the big introspectable difference the Martians report, it does not follow that Martian color-words are ambiguous. Color can be regarded as doubly determinable, by x-ing and by y-ing, for the Martians, just as shape is doubly determinable, by sight and by touch, for us. Grice goes on to question this comparison of shape and color. It is quite conceivable, he claims, that the correlation between x-ing and y-ing a color should break down in a limited class of

16. "Some Remarks About the Senses," *Analytical Philosophy*, first series, edited by R. J. Butler (Oxford: Basil Blackwell, 1962), pp. 133-53.

17. Much of my discussion of Reid goes against this claim. It is the sort of information received, rather than the experiences which accompany its reception, which distinguish the senses. For an examination of Grice's central argument for his claim, see J. W. Roxbee Cox, "Distinguishing the Senses," *Mind*, 79 (1970): pp. 530-50; and for a suggested emendation of Roxbee Cox's account, see my "The Primary Objects of Perception," *Mind*, 83 (1976): pp. 189-208, especially p. 196.

cases. In these cases, "the conflict would render decision about the real color of the objects in question impossible" (p. 148). He is inclined to think, on the other hand, "that a corresponding limited breakdown in the correlation between sight and touch with regard to shape is not conceivable" (p. 149). We cannot coherently suppose both

(a) that, in a world which in general exhibits the normal correlation between sight and touch, some isolated object should standardly feel round but standardly look square, and also (b) that it should be undecidable, as regards that object, whether preference should be given to the deliverance of sight or to that of touch. (p. 149)

I take it that satisfaction of condition (b) does not imply that it is impossible to render a decision about the real shapes of the divergent object. A divergent object would not have just one real shape, but two, one perceptible by touch and one perceptible by sight.¹⁸

Grice considers two sorts of cases, *Case A* and *Case B*, to show that conditions (a) and (b) quoted above cannot be jointly satisfied. In *Case A*, an apparent limited breakdown between tangible and visible shape is not really a breakdown. In *Case B*, an apparent limited breakdown becomes unlimited. Suppose, as is allowed in *Case B*, that a divergent object has the power to infect the spatial properties of nearby objects. If I cannot determine the real shape of the object by handling because my finger looks to be tracing a spatial path different from the spatial path it feels to be tracing, I can use other tests. An object which passes through a square hole of a certain size, but not through any holes of a smaller size, is really square, no matter how it looks. If an object tactually passes through a square hole while visually remaining on the other side, this spatial separation of tangible and visual position cannot be confined to the divergent

18. Similarly, a color-divergent object could be regarded as having two colors, one detected by x-ing (or by females, by the right eye, and so on), the other detected by y-ing (or by males, by the left eye, and so forth). If it is undecidable whether preference should be given to one or the other, rather than saying that nothing can be decided about the real color of the object, we can abandon or modify the principle that if something really has one uniform color then it cannot have another.

object. If such a separation occurs as you push a tactual object through a hole, for example, your finger tangibly moves through the hole while visually remaining on the other side. In this way, the hypothesis of limited breakdown leads to unlimited breakdown as the separability of tangible and visual location is transmitted from one object to another.¹⁹

After the forthcoming discussion of impenetrability, I shall attempt to show that we can coherently suppose conditions (a) and (b) to be jointly satisfied. This attempt will involve a case much closer to *A* than to *B*. In this paragraph I shall defend Grice's treatment of *Case A*, where the divergent object does not have the power to infect the spatial properties of nearby objects. In this case, Grice argues, the only real shape of the divergent object is determined by touch. Either the divergent object has invisible parts, or its visual outline somehow extends misleadingly beyond its genuine tangible outline. When Grice imagines that a finger is seen to cut through the corner of the visible square of an object that feels round, he says that "such a lack of 'visual solidity' would be enough to make us say that the object is really round, in spite of its visual appearance" (p. 149). If, on the other hand, we try to suppose that the visible outline of a divergent object does possess visual solidity with respect to our fingers, and we retain the *Case A* hypothesis, the visible outline will also be a tangible outline. Suppose that at least part of the visual boundaries of a divergent object extend beyond its tactual boundaries and that our fingers are seen to be in contact with these parts of the visual boundary. What happens when the fingers apply some pressure? If the visual boundary is visually solid with respect to our fingers, our fingers will look not to penetrate the visual boundary. Unless the correlation between where the fingers look to be and where they feel to be is upset by contact with the divergent object, contrary to the *Case A* hypothesis, the fingers will then also feel not to be penetrating the visual boundary. But a boundary which our fingers can touch but cannot penetrate while they retain their normal tactual sensitivity is a tactual boundary. The supposition that the visual

19. Thomson, acknowledging Grice's influence, further develops this ingenious line of argument on pp. 640-45.

boundaries extend beyond the tactual boundaries is thus reduced to absurdity.

Solidity is understood here to be the property of impenetrability. Fingers are paradigms of solids which are both visually and tangibly impenetrable. The argument above shows that if we can coherently suppose that there are visual solids which are not also tangible solids, then the visual solidity should not be tested by visual impenetrability to fingers. But I think there are other possible tests of visual impenetrability. In order to imagine divergent objects of the sort Grice describes, we should investigate the logic of impenetrability.

I have elsewhere questioned the widely received view that no two material objects can be in the same place at the same time.²⁰ It is logically possible, I claim, that two billiard balls should roll towards and pass through one another. The following line of argument against this claim has recently been entertained by Harold W. Noonan:

. . . as the billiard balls begin to merge there ceases to be even one billiard ball there. For even one billiard ball to be there, it may be said, there has to be a spherical lump of the stuff billiard balls are made from, of a certain size, *surrounded by a substance of a distinct kind*: thus the normal situation of a billiard ball on a billiard table.²¹

20. "Locke, Leibniz, and Wiggins on Being in the Same Place at the Same Time," *Philosophical Review*, 79 (1970): pp. 75-82. The view I oppose has been held by so many that it is a comfort to keep discovering evidence that it has not been held by everyone. Here is a passage from the article "Atom" by James Clerk Maxwell in the *Encyclopedia Britannica*, 9th edition (1875-89), quoted in the article "Molecule" by James Hopwood Jeans in the 11th edition.

Boscovich himself, in order to obviate the possibility of two atoms ever being in the same place, asserts that the ultimate force is a repulsion which increases without limit as the distance diminishes without limit, so that the two atoms can never coincide. But this seems an unwarrantable concession to the vulgar opinion that two bodies cannot co-exist in the same place. This opinion is deduced from our experience of the behavior of bodies of sensible size, but we have no experimental evidence that two atoms may not sometimes coincide.

21. "Can One Thing Become Two?" *Philosophical Studies*, 33 (1978): pp. 205-6. Noonan himself says that the objection is superficial although he does not indicate how he thinks it should be handled.

If being surrounded by a substance of a distinct kind is a necessary condition of the existence of a billiard ball, it is a necessary condition of the existence of any material object. There is presumably nothing special in this respect about spheres or about objects made of substance of the kind of which billiard balls are made. So there are reasons independent of the logical possibility of place-sharing to reject the suggested necessary condition of existence. If sixteen cubes of ivory, or of the same metal, are stacked together to form a larger cube, the smaller cubes in the middle continue to exist even though they are surrounded by a substance of the same kind. There may in fact be a thin layer of some other substance such as air or water or oil between the stacked cubes, but there is no logical incoherence in the supposition that two surfaces of independently movable objects have nothing between them. Even if the cubes should stick together, and not be independently movable, it is not obvious that they thereby cease to be cubes. Would a plaster cast of a figure of Hercules cease to exist merely by being encased in more plaster even if no air or anything else is trapped between the new plaster and the old, and the new plaster after it sets sticks as tightly to the old plaster as it coheres with itself? If the surrounding plaster is pink and the figure of Hercules is white, we can recover the bust by carefully chipping away pink plaster until we come to white. (And if difference of color is sufficient for distinctness of kind, the principle Noonan considers is consistent with two billiard balls of different colors passing through each other.²² If the surrounding pink plaster gradually fades, and there is eventually no way to distinguish the new plaster from the old, then the figure may be unrecoverable. The whole hunk

22. What if two balls of different colors exactly coincide? A ball that looks just like an ordinary white ball will not appear to occupy just the same space as a ball that looks just like an ordinary red ball appears to occupy. There is a lot of latitude in filling in the details of stories concerning *Interpenetration*. We can suppose, for example, that when a surface that absorbs n percent of incident light of a certain wavelength coincides with a surface that absorbs m percent of incident light of that wavelength, the two coincident surfaces together will absorb either $n + m$ percent or 100 percent of incident light of that wavelength, whichever is less. On this supposition, some exactly coinciding objects would appear a deeper black than black velvet.

of plaster will then be intrinsically indistinguishable from another uniform hunk of plaster of the same size. The second hunk does not contain a single figure of Hercules any more than it contains innumerable figures of Hercules and other characters. If the first hunk continues to contain a single figure of Hercules, the difference between the hunks is not a difference of intrinsic qualities. Consider a two-dimensional analogue. On Tuesday you paint a square shape on a wall. On Wednesday you paint the rest of the wall the same color. Your edging technique is so perfect that, without painting over any of the square you painted on Tuesday, you leave the wall with a perfectly uniform coat of paint. The boundary between the area painted on Tuesday and the area painted on Wednesday is undetectable. Still, when you look at the painted wall, part of the area you see was painted on Tuesday, and that part is square. Similarly, part of the first hunk of plaster was set before additional plaster was added, and that part is a figure of Hercules. It persists even while surrounded by a substance indistinguishable from the substance of which it is made.

It does, ironically, suit my purposes to stipulate a meaning of *material object of the same kind* which renders Noonan's principle irrefutable by place-sharing. Let us say that two material objects are of the same kind if and only if they exclude each other and each excludes anything the other excludes. It is then true by the definition of 'same kind' that no two material things of the same kind can occupy the same place at the same time. Two material things are to be counted as of different kinds if they can occupy the same place at the same time or if there is any third thing which can share the place of one but cannot share the place of the other. If there are two kinds of objects, either any object or no object of one kind will exclude any object of the other kind.

Entertain for a while the following fantasy: although rocks from the Moon exclude each other and terrestrial objects, and rocks from Mars also exclude each other and terrestrial objects, Mars rocks and Moon rocks pass through each other without difficulty. A sculpture carved from Moon rock when placed directly on a pedestal carved from Mars rock will drop right through it to the floor. A Mars rock pedestal can be used for a

Moon rock sculpture only if a layer of some third kind of substance is interposed.

If one object cannot at the same time occupy the place occupied by another object, let us say that they are directly related by exclusion. And let us say that two objects related by the ancestral of direct exclusion belong to the same exclusion family. Since Moon rock objects and Mars rock objects belong to the same exclusion family, their sizes and shapes can be compared. Objects of each kind can coincide with terrestrial objects or exactly fit through holes cut in terrestrial objects.

Now let us reconsider Grice's *Case A* divergent object whose visual outline extends beyond its tangible outline. Our fingers and whatever is bounded by the visual outline are not directly related by exclusion, but they may belong to the same exclusion family. It may be, for example, that the visible outline will fit through a square hole of a certain size cut in Mars rock but not through any smaller hole cut in Mars rock. If the tangible outline of a divergent object extends beyond its visible outline, it may still be that the visible outline will exactly fit a square hole cut in Mars rock while the round tangible outline passes right through the Mars rock surrounding the hole. If the supposition that there could be different kinds of objects, as I am here understanding 'different kinds', is coherent, then so is the supposition that there should be isolated divergent objects. The matching and fitting tests for shape involving objects of one kind need not agree with similar tests involving objects of another kind.

If an object is tightly wrapped with a relatively thin material, the shape of the resulting package is approximately the shape of the original object. Thus, as Thomson remarks, the shape of an invisible object could be made visible by wrapping it in sheets of a visible material (p. 640). Similarly, the shape of an intangible object, an object our fingers can pass right through, could be made tangible by wrapping it in sheets of a tangible material. As in any case of wrapping, the wrapping material and the object wrapped must exclude each other. If appropriate wrapping materials were available, then, the tangible shape of a divergent object could be made visible by wrapping it with one sort of material, and the visible shape of the object could be made tangible by wrapping it with another sort of material.

Although I have assumed that it is logically possible for two objects to occupy the same place at the same time in my attempt to provide a coherent description of a divergent object, a divergent object need not to be regarded as a composite of an object of one kind which occupies some or all of the space occupied by a differently shaped object of a different kind. We may suppose, if we like, that the visible and tangible boundaries of a divergent object are not only inseparable but also could not have been separate. If two billiard balls wholly or partly interpenetrate, they might become inseparable in the sense that nothing can be done to separate them. Still, something could have been done to ensure that they would have been separate now: they could have been kept apart. The boundaries of a divergent object, we may suppose, are not like this.

Even if it is granted that a divergent object is a single object rather than a composite of interpenetrating objects with different shapes, one may resist the contention that a divergent object has two shapes. After all, "No object can have two different shapes at the same time," like "No two objects can be in the same place at the same time," may belong to a stock of examples one is accustomed to regard as necessary truths. "Why not," it can be asked, "regard a divergent object as having a single shape determined by whichever boundary, tangible or visible, extends farther in a given direction?" I see nothing to favor this suggestion over the similar, but incompatible, suggestion that a divergent object be regarded as having a single shape determined by whichever boundary, tangible or visible, extends less far in a given direction. I reject both suggestions because I take the fitting and matching tests as definitive of a thing's shape, not merely as normally reliable evidence of a thing's shape that would become unreliable in the circumstances I have described. In our world, so far as we know, an object that passes any fitting and matching test for being of a certain size and shape can pass any other such test. I have tried to show that passing one test for being of a certain shape is logically compatible with passing another test for being of a different shape. I have not considered the case where there is just one anomalous test that disagrees with all the others. A divergent object of the sort I have described that passes one fitting and matching test for being of a

certain shape, so long as it persists unchanged, can pass indefinitely many such tests. It can also pass indefinitely many tests for being of a different certain shape. Whether it passes a test for being of one shape or another depends on which sort of object it is tested against. According to the fitting and matching tests, a divergent object can be both a cube and a sphere.

Many further possibilities of objects not directly related by exclusion can be described. I shall briefly sketch three. (1) If an invisible object excludes only other invisible objects but does exclude (invisible) objects of kind *K* which in turn exclude visible objects of kind *J*, two layers of wrapping, such as an inner layer of kind *K* and an outer layer of kind *J*, are required to render its shape visible by the wrapping technique. Another invisible object might require even more layers of wrapping. Similar complications can be described for tests of fitting and matching. (2) The parts of a creature's body need not be all of the same kind even if, what is also imaginably otherwise, every part of the body is directly related by exclusion to every other part. A divergent object could thus have two tangible shapes, one perceived by the right hand and one perceived by the left, if the left and right hands were of different kinds. (3) The boundaries marked by differences in olfactory or auditory qualities described in the discussion of Reid could be the boundaries of objects which, although not directly related by exclusion to any parts of the bodies of the creatures who perceive them, do belong to the same exclusion family.

If we can coherently suppose that there should be divergent objects, how does this affect Molyneux's question? If a divergent object which is tangibly a cube but visually a sphere and a divergent object which is tangibly a sphere but visually a cube are shown to the man born blind now made to see, what question should he be asked? Not "Can you tell just by looking which object is tangibly a cube and which tangibly a sphere?" Given that he does not know whether or not the objects before him are divergent, he can only guess; but this does not distinguish him from normal sighted persons, who also can only guess in this circumstance. The question to ask is rather "Can you tell just by looking which object is visually a cube and which visually a sphere?" It does not matter to the question which concerns

Molyneux most whether the objects whose shapes are perceived by touch have coincident visual boundaries, divergent visual boundaries, or no visual boundaries. (In this last case, the man born blind can be shown a different set of objects.) Molyneux's question is not primarily about the visual reidentification of objects previously perceived by touch, but about the visual identification of shapes previously perceived by touch.

Although my flights of fancy, if coherent, show that there might be cubes very different from the cubes we know, they also help to show that 'cube' is not ambiguous between the different kinds of cubes there are or might conceivably be. The contention of Berkeley's we began with, that there are, strictly speaking, no common sensibles, has been subverted in many of my descriptions of outlandish possibilities. The boundaries of divergent objects, the boundaries of objects only remotely related by exclusion to our own bodies, and the boundaries marked by sensible qualities which are neither visible nor tangible have all been described as spatially related to each other. Shapes of these different kinds of boundaries are not different kinds of shapes; they can be spatially congruent. If one perceives with one sense the shape of a boundary which is congruent with a boundary whose shape one perceives with another sense, then one perceives the same shape with two senses. A divergent object is divergent in having two shapes, not in having one shape which looks different from the way objects standardly look which standardly feel to have the shape the divergent object feels to have. The divergence of visible and tangible shape is due to the spatial relations between them, not to their violating a general regularity. The congruence of visible and tangible shape, similarly, is not a matter of their instantiating a general regularity.²³

23. I am grateful to the editors for their comments on an earlier draft of this paper. I suspect that each remains unpersuaded by some of my suggestions.

Abstraction Reconsidered

PETER GEACH

As I return to a new treatment of this subject after many years, I am conscious of a methodological difficulty of which I was then strangely unaware. I wrote as if abstractionism were a *thesis* about the nature and formation of concepts, one that could be definitely stated and definitively refuted. This now strikes me as a mistake: I regard abstractionism not as a thesis to be refuted but as a confusion to be dispelled; in abstractionism we have not a picture that can be closely inspected and criticized, but a mirage that disappears when closer inspection is attempted. It is strange to me now that I did not think of this at the time; for the Wittgensteinian principle that what one judges to be so, though it may be false, cannot be nonsense, was explicitly in my mind when I wrote *Mental Acts*, and I was also aware that many philosophical 'theses' are arguably not false but nonsensical. A teacher of philosophy is often engaged in practical criticism of 'theses' that he must regard as simply muddles; he must try to convey what in his view the muddle is, without either falling into muddle himself or misrepresenting the muddle by turning it into a plain falsehood. I have no general idea of what one should do in this predicament; I do my best from case to case, and that is what I shall try to do here.

Ascription of Concepts and Judgments

Abstractionism is on the face of it a theory about the formation of concepts: the word 'concept' being here used, not in the

objective sense in which Frege used '*Begriff*', but as relating to the abilities of an individual. But here our troubles start: so understood, the concept *concept*, as Wittgenstein said, is a vague one; it is hard to delimit our ways of talking about concepts in general, or about particular concepts, e.g. the concept *cat*. The source of difficulty, however, is not that in such cases we use the definite article: there is no more a real problem how 'the' concept *cat* can be shared by many people than how 'the' ability to swim can be.

Concepts, being located in the realm of human abilities or capacities, are to be understood by considering the acts in which they are characteristically exercised; and I still think a good sense can be given to the old dictum that concepts are exercised in acts of *judgment*. There has indeed been a lot of bad philosophy written using this word (or the corresponding French or German word); but I shall not on that account replace it with another, for I know no better. Again, there has been much confusion about the relation between the state of mind called belief and the act of judgment: belief is another topic, and a separate one, which I shall avoid. A man often finds himself confronted with a question or problem, to which he must then and there think up an answer: his doing so is an act of judgment. This is an illustration, not a definition; but I think it will suffice for my purposes; there will be no need to consider whether mental acts widely diverging from these typical cases should still be regarded as acts of judgment.

In *Mental Acts* I brusquely dismissed the possibility of ascribing acts of judgment, or the exercise of concepts, to brute animals: roughly, because their pattern of life differs so much from that of human beings as to make such ascription pointless. It now seems to me that here I went about things too fast; there was, however, a clear justification for *concentrating* on human examples. In so far as we are justified in applying psychological terms to brutes, this comes about because we rightly see some analogy between the patterns of living in them and in human beings; and of course it is the human side of the comparison that we understand better. So, as a matter of method, no harm is done by concentrating on human examples; when we have achieved an understanding of judgments and concepts and con-

cept-formation in these cases, then we may or may not find sufficient similarity in some aspects of a brute's life to ascribe judgments and concepts and concept-formation to the brute.

A similar question of method is: should we concentrate on judgments expressed in language? Here again there was good reason for doing so. Putting a judgment into words is certainly not the only way of expressing it; on the other hand, there are recognizable kinds of utterance, e.g. declarative sentences, that are characteristic ways of expressing judgments. If we are to ascribe a judgment to a man on the ground of other behavior, this involves a complex appraisal of his wants, purposes, and intentions. But a native speaker of a language will certainly most often use it with understanding; and in the most corrupt society lying is necessarily the exception, or communication would be impossible; so a man's words will often justify us very directly in ascribing to him a certain judgment. Even if a man is lying, he will be exercising the same concepts as if he had *not* been lying; so we can ascribe concepts on the basis of linguistic behavior even more securely than judgments.

It is only this point of method that I was making in *Mental Acts*; I was certainly not *defining* the possession of concepts solely via linguistic behavior and the related abilities. I expressly gave what would serve as counterexamples to such definitions: e.g. victims of aphasia who can still play such games as bridge or chess, and must therefore still have the corresponding concepts. I might similarly have instanced the manifest mastery of sophisticated concepts by adult deaf-mutes who have never learned any word-language. The concentration on linguistic expressions was justified, not because here *alone* may concepts and judgments be ascribed, but because here the ascription is *easier* and *safer*. There is of course the opposite danger of an unbalanced diet of examples; I can only say that I do not forget this danger, and strive not to be misled.

I suppose these days I may be expected to say something about the ascription of concepts and judgments to brutes and to automatic machines on the ground of *their* linguistic behavior. About the first I remain sceptical. Years ago I heard a psychologist who gave lectures in Oxford telling a story of an ape that could *speak* like a young child; I told him he could tell

that to the Marines, and I find no reason to regret that I did so. This story does not now go the rounds; instead we are told of apes who have mastered languages in non-vocal media—in gesture, or in visible symbols. But when I consider how the alleged symbolic performances of certain apes compare with the actual performances of some Leeds students beginning to learn logic, I am sceptical about these stories too; and for now it must suffice that I say this.

As for machines, it is quite certain that they make no judgments and form no concepts; for these activities are part of the life of a living being, and by all sorts of criteria these machines are not alive. We are familiar with the transferred use of language whereby a book, or even a theory found in books, is said to maintain one thing, or argue a second thing from certain premises, or fall into inconsistency—predications primarily made of live human beings. We are not misled here; it shows a certain wilfulness, a certain appetite for self-deception, when people insist that similar transferred uses of language must be taken literally when we are talking about computers. But suppose we could build a human being piece by piece as we build a computer? But you cannot, you know; so if you please we won't suppose it.

The Individuation of Concepts

It sometimes appears natural to compare a man's set of concepts with a carpenter's set of tools: all the more because items in the tool-kit—the hammer, the ruler, the glue-pot—are manifestly heterogeneous. But in one way the comparison limps badly: the tools could be listed, and there would be just so many of them; there could be no question of an exact list of a man's concepts. In *Mental Acts* I alluded to the abilities exercised in playing chess. These abilities are distinguishable, and in some measure independent: chess could be taught by the teacher's first playing a simplified game, say with only kings and pawns, and gradually introducing more pieces in successive games. But it would not be sensible to ask just how many abilities are involved in playing chess. The like holds with concepts exercised linguistically.

Whether we say two people have the same concept or different concepts, in describing a case where their abilities overlap but do not coincide, need not be very important, so long as we are clear what the two severally can and cannot do. A man blind from birth can to some extent discourse intelligently about colors; he cannot relate this discourse to visual experience, because he has none. In saying his concepts of colors are different from ours, we are just re-stating and re-emphasizing this disability, which is a *logical consequence of his blindness*; there is no foundation here for any *empirical* theory of concept formation. If as in *Gulliver's Travels* we heard a story of a blind instructor training his pupils to recognize the colors of pigments by feeling and smelling, then initial incredulity would be in order; but the supposed feat is no more inexplicable than some quite common discriminative powers and no more incredible than some rare discriminative powers that can be demonstrated. There would be little reason to deny color concepts to such blind men. And the connection of a concept with *any* power of sensory discrimination may be unimportant. Cavendish's concept of electric current tied in with direct experiences of electric shocks; he acquired surprising skill in the use of his own body as an electrical instrument.¹ But this does not make an important difference between his concept of electric current and some other early nineteenth-century researcher's concept, unconnected with discriminations of electric shocks.

Abstractionism Expounded

For abstractionists, concepts are essentially capacities for recognizing recurrent features in the world. We learn to attend to such a feature and ignore other features presented simultaneously. But there is a difficulty in expounding their view on the crucial question how concepts are exercised in acts of judgment. Let us take a judgment expressed in words, and let us suppose that for each word there is a recurrent feature of reality that we have learned to recognize. When we make a judgment, we may not be confronted with a situation in which all of the features

1. See A. J. Berry, *Henry Cavendish* (London: Hutchinson, 1960) pp. 181-89.

corresponding to the words are discernible: I may judge that cats eat mice in a situation where there is no cat, no eating, and no mouse. If the concepts *cat*, *eating*, and *mouse* were nothing but recognitional capacities, then it would be unintelligible that these capacities should be exercised in the judgment.

Understanding a sentence like 'Cats eat mice' may indeed be said to depend on grasping what each single word means; and it is this grasp that abstractionists believe to involve the relevant recognitional capacity. Let us begin by emphasizing the actual dependence of sentence meaning on word meanings. We do indeed learn words for the most part *in* sentences; I shall discuss later the learning of single words by ostensive definition. But having learned the meaning of words, we go on to understand, and also to produce, sentences that we have never previously learned.

What remains obscure is *how* the understanding of words, if it is itself a matter of recognitional capacities, can yield the understanding of a new sentence that answers to an observed or visaged state of affairs. I can only suppose that abstractionists are 'held captive by a picture', by an inappropriate metaphor, which at first seems to work, but dissolves finally into nonsense.

The word 'feature' itself suggests how this might happen. Observing faces, we can notice and reidentify certain recurrent features: a Roman nose, a Habsburg jaw, and so on. Putting these features together, as in a police identikit, we can reproduce a face we have seen, or imagine one we have not seen. Similarly, it might be supposed, we pick out recurrent features of states of affairs, and are thus enabled to describe new observed states of affairs, or imagine new unobserved ones.

There are two ways in which this model breaks down. First, a feature of a face can in general be recognized and observed all by itself, apart from the rest of the face; and abstractionists clearly suppose that the like can be done with the features of states of affairs, to which the separate words of sentences correspond. That is why they lay such stress on the learning of separate words by ostensive definition. But many words are in principle unlearnable this way. This is conspicuously true for logical words—though, as we shall see, abstractionist accounts have been attempted even for these. But there are many other words whose

meanings could not possibly be first learned in isolation and then deployed in a sentence. A number word, or again a word like 'big' or 'small', can be applied only in relation to some kind of things: abstracted from the kind of things whose number or normal size is in question, the term has nothing to latch onto.

But there is a second fatal difficulty that would arise even if the features of things in general were things we could learn to recognize separately like the features of faces: there could be no general technique, like the police identikit procedure, for fitting together features of reality into a state of affairs. Human faces resemble one another in a way Humpty Dumpty found confusing: two eyes horizontally set apart at the top, a nose going down vertically between them, a mouth under that . . . There is no such uniform organization of features of reality in states of affairs, nor even a limited number of types of organization—just as there is no one common structure of all sentences in a language, nor even a limited number of sentence-structures. Abstractionist thinking would suit well with the bad old logic, for which there were just four forms of simple categorical proposition and a few odd forms of complex propositions like hypotheticals and disjunctives. But there are countless forms of sentences, countless varieties of states of affairs. Logicians and linguists have indeed brought much order into the analysis of sentences by recursive procedures that can generate, and enable us to describe, infinitely many forms; but no such procedures are known that will generate all the sentences of a natural language—yet such a language can be learned.

Reverting to ostensive definition, we may notice that the idea many people have of how language is learned is largely mythical. They imagine a child trained to imitate its parents' utterances of single words in the presence of the appropriate *kind of thing, and thus learning simultaneously to pick out some feature in the world and the word for that feature*. But even with affectionate and leisured parents such training can constitute only a minuscule part of the process by which language is transmitted. In any case it could afford no explanation of how sentences are understood.

In his *Thinking and Experience*, to my mind the best exposition and defence of abstractionism, H. H. Price gives a far

better account of language learning: we learn that a word answers to a recurrent feature because we repeatedly observe the use of the word in sentences that answer to situations in which the feature is prominent. Let us spell this out in an example of a child growing up and learning language in a harassed family who have no leisure for careful training of the child in the language, item by item. Father comes home in an ill temper after an unlucky day at the dog-races; he trips over the cat as he comes into the house, and furiously kicks the animal into the street. The cat is dramatically prominent in the situation; the word 'cat' (*inter alia*) may be prominent in Father's language. Price would have it so that by concurrent abstraction of cat-hood from situations involving cats and of the word 'cat' from sentences containing it we acquire the concept *cat* and the meaning of the word 'cat'.

Clearly this is much better than the story of words being learned by ostensive definition. However, both stories have a common defect. We do not understand a word *at all* unless we assign it to the right category; of course this does not mean we need possess the concept *category*, only that we must rightly locate the word in our linguistic practice. Ostensive definition clearly gives no account of the assignment of words to categories: if even an adult user of language is being taught a new word by ostensive definition, he will grotesquely misunderstand if he gets the category wrong. Price's account, by bringing in sentences, is clearly superior, but even this does not explain our learning coherent sentences rather than strings of words severally associated with features of a situation. A sentence is not a word salad.

Logical Concepts

I shall here be concerned not with the concepts involved in intelligently using the terms of art in a treatise of logic, but with those supposed to be involved in intelligently using such particles as 'all', 'not', and 'or'. Two ways of treatment are to be found in writers of an abstractionist cast of mind: they may be called the objective way and the subjective way. By the objective way of attacking the problem, we abstract a certain feature to be

found repetitively in situations whose resemblance is shown by our using some one particle in all the sentences describing the situation, and learn to use a logical particle like 'all' or 'not' or 'or' to express the presence of this feature. The subjective way leads us rather to concentrate on introspectible features of our own reactions to situations.

There is very little, I think, to be said in favor of the objective way. Abstractionists who try to go this way sometimes regale us with imaginary stories of nursery crises, for whose description an adult would use a sentence containing 'some' or 'not': how comes an adult to do this, unless he has in infancy noticed the somehow or nottishness involved in this sort of eventuality? This sort of plea is quite vain until the abstractionist can give a logically coherent account of the features that are supposed to be reached by discriminative attention, corresponding to the words 'some' and 'not'. We might have to wait a long time for such an account to be even attempted.

This argument from silence may be supplemented by positive considerations: we have definite reason to dismiss the idea that as someone has a concept *man* or *red*, so he has a concept *some* or *all* or *not* or *or* or *if*. Let us begin with negation. Even at the level of mere sorting, a concept is essentially two-sided: sorting the *As* from the non-*As* is the same as sorting the non-*As* from the *As*. Thus it is clearly wrong to think that the concept of an *A* needs some supplementation with a concept *not*, in order for a man to know what it is *not* to be *A*.

Similarly, it is not clear how objective alternativeness could be found in a situation that I describe by using the word 'or'. How we learn the use of this word is rather mysterious. I have read a philosopher saying that we learn to assert 'P or Q' in circumstances where it would be justifiable to assert P or justifiable to assert Q. This appears very unlikely: why should we then not straightforwardly come out with P, or else come out with Q, instead of the more complicated and less informative utterance 'P or Q'? In fact, we have e.g. disjunctive memories like 'I left my spectacle-case in the drawing-room or else in the study', when neither disjunct is given in memory; but supposing the memory to be correct, there is no discernible alternativeness as a feature of the situation remembered.

Failing to find features of the world observed to associate with the logical words, abstractionist thinkers have sometimes tried to associate features of inner experience: self-inhibition for 'not', choice for 'or', and so on. No doubt such experiences sometimes do accompany the use of such words; but they do not by any means always or mostly accompany the current use of the words, nor is communication of such experiences at all important for getting over a message by sentences in which such words occur.

A prudent abstractionist, I think, would just not try to apply his theory to logical words.

Relational Concepts

For words of a relational meaning, the idea of our getting at that meaning by concentrating our attention on some recurrent feature of reality and ignoring other features simultaneously given is one that breaks down almost at once. I argued this point in *Mental Acts* for the concepts *big* and *small*. An abstractionist critic of mine replied that it is quite easy to understand our noticing this common feature, *large size given the sort of thing*, as one shared say by a mouse and an elephant. But what is *the sort of thing* in question? Minnie is at once a mouse, a rodent, and an animal; she is, let us suppose, a big mouse, but a small rodent and a small animal. If we abstract from the sort of thing indicated by the word 'mouse' or 'rodent' or 'animal', then Minnie cannot be considered as big or small. Discriminative attention to bigness is an absurdity.

Similarly for the sort of concept that answers to a transitive verb. The striking of blows is a recurrent feature of reality if anything is; blows in a fight can even be counted. But a man's concept of *striking* is extremely defective if he cannot answer the Leninesque question 'Who whom?' To understand the verb 'to strike' one must know the difference between 'A struck B' and 'B struck A'; but if we abstractively erase the subject and the object, the difference becomes invisible. To put it another way: *striking* and *being struck* are different, but they necessarily go together both in events and in our thoughts. Only in the context of a judgment can *striking* and *being struck* be distin-

guished; and in the act of judgment we are thinking of both together, though not of both interchangeably. If the sword of abstraction could dis sever them, then (to borrow a figure from Frege) it could scarcely be a magic sword that would also reunite them; and if it were, what could give us the skill to reunite in the right way 'A struck B; B was struck by A' rather than in the wrong way 'B struck A; A was struck by B'? Certainly not the abstraction that makes us ignore and forget about A and B.

Numerical Concepts

Some critics of mine have stumbled at the phrase 'abstract counting'. The verb 'to count' has one standard application to the procedure of producing the successive numerals of a given language, as opposed to counting objects of some sort; this is what I call abstract counting, and I do not care a rap whether the phrase is standard English. Abstract counting can and must be learned independently of the use of numerals to count objects. A normal English child could count on accurately from 'two hundred and fifty-seven' till told to stop; no actual count of objects, 256 of which have already been counted, need have been performed even once in the child's life.

The series of numerals is generated according to a rule. Given a finite set of numerals, we can continue abstract counting indefinitely. (In English the required set is: the numerals from 'one' up to 'twenty'; 'thirty', 'forty', etc. up to 'ninety'; 'hundred'; 'thousand'. We may ignore baroque extravagances like 'billion' and 'quadrillion': the Biblical style 'ten thousand thousand' is obviously all that we need in theory.) The rule for doing this is of course not set forth in words, though it would not be very difficult to do so; but it is well grasped in practice. If the rule were set forth in words, theoretically there could be disputes about its interpretation; in fact, even without the rule, there are no disputes as to what is correct practice.

Abstract counting has sometimes been called a parrotlike procedure. A parrot might well learn to recite some stretch of the numeral series; knowing how to go on indefinitely is a human prerogative. The key ability is the ability to notice how

Failing to find features of the world observed to associate with the logical words, abstractionist thinkers have sometimes tried to associate features of inner experience: self-inhibition for 'not', choice for 'or', and so on. No doubt such experiences sometimes do accompany the use of such words; but they do not by any means always or mostly accompany the current use of the words, nor is communication of such experiences at all important for getting over a message by sentences in which such words occur.

A prudent abstractionist, I think, would just not try to apply his theory to logical words.

Relational Concepts

For words of a relational meaning, the idea of our getting at that meaning by concentrating our attention on some recurrent feature of reality and ignoring other features simultaneously given is one that breaks down almost at once. I argued this point in *Mental Acts* for the concepts *big* and *small*. An abstractionist critic of mine replied that it is quite easy to understand our noticing this common feature, *large size given the sort of thing*, as one shared say by a mouse and an elephant. But what is *the sort of thing* in question? Minnie is at once a mouse, a rodent, and an animal; she is, let us suppose, a big mouse, but a small rodent and a small animal. If we abstract from the sort of thing indicated by the word 'mouse' or 'rodent' or 'animal', then Minnie cannot be considered as big or small. Discriminative attention to bigness is an absurdity.

Similarly for the sort of concept that answers to a transitive verb. The striking of blows is a recurrent feature of reality if anything is; blows in a fight can even be counted. But a man's concept of *striking* is extremely defective if he cannot answer the Leninesque question 'Who whom?' To understand the verb 'to strike' one must know the difference between 'A struck B' and 'B struck A'; but if we abstractively erase the subject and the object, the difference becomes invisible. To put it another way: *striking* and *being struck* are different, but they necessarily go together both in events and in our thoughts. Only in the context of a judgment can *striking* and *being struck* be distin-

guished; and in the act of judgment we are thinking of both together, though not of both interchangeably. If the sword of abstraction could dissever them, then (to borrow a figure from Frege) it could scarcely be a magic sword that would also reunite them; and if it were, what could give us the skill to reunite in the right way 'A struck B; B was struck by A' rather than in the wrong way 'B struck A; A was struck by B'? Certainly not the abstraction that makes us ignore and forget about A and B.

Numerical Concepts

Some critics of mine have stumbled at the phrase 'abstract counting'. The verb 'to count' has *one* standard application to the procedure of producing the successive numerals of a given language, as opposed to counting objects of some sort; this is what I call abstract counting, and I do not care a rap whether the phrase is standard English. Abstract counting can and must be learned independently of the use of numerals to count objects. A normal English child could count on accurately from 'two hundred and fifty-seven' till told to stop; no actual count of objects, 256 of which have already been counted, need have been performed even once in the child's life.

The series of numerals is generated according to a rule. Given a finite set of numerals, we can continue abstract counting indefinitely. (In English the required set is: the numerals from 'one' up to 'twenty'; 'thirty', 'forty', etc. up to 'ninety'; 'hundred'; 'thousand'. We may ignore baroque extravagances like 'billion' and 'quadrillion': the Biblical style 'ten thousand thousand' is obviously all that we need in theory.) The rule for doing this is of course not set forth in words, though it would not be very difficult to do so; but it is well grasped in practice. If the rule were set forth in words, theoretically there could be disputes about its interpretation; in fact, even without the rule, there are no disputes as to what is correct practice.

Abstract counting has sometimes been called a parrotlike procedure. A parrot might well learn to recite some stretch of the numeral series; knowing how to go on indefinitely is a human prerogative. The key ability is the ability to notice how

many times some operation has been performed; having counted several times from 'one' to 'ten', we can notice how many times we have so counted, and then how many times we have counted *ten* counts from 'one' to 'ten'—and so on, and so on, *indefinitely*. This sort of pattern is a free creation of the human mind; no pattern in the natural world is there to suggest it.

In abstract counting, then, what is fundamental is the use of a numeral as (in Wittgenstein's phrase) the exponent of an operation. Counting objects of a kind adds two things to abstract counting: first, the recognition that the objects are all of one kind (as Frege puts it, all fall under the same *Begriff*); second, the setting up of a one-to-one correlation between these objects and the numerals from 'one' up to some other numeral, which is then taken to give the number of objects of that kind.

Abstractionists give a very different account of numerals. Numerals are supposed to answer to recognizable common features of sets or groups: a group may consist of similar objects, or of any old objects (as in some recent set-theoretical presentations of elementary arithmetic) without their needing to be specially similar, or again of 'abstract units' (a very mirky notion). I shall not spell out the refutation of these lines of thought: it has been given once for all in the works of Frege, particularly his *Grundlagen*.

Once we have the key idea of a numeral as the exponent of an operation, it is easy to see how the various elementary operations of arithmetic are to be regarded: multiplication is addition repeated so many times, and division is subtraction repeated so many times; exponentiation is multiplication of 1 by some number so many times over. Addition and subtraction themselves are in principle similarly explicable, in terms of how many times the procedure of passing from a number to its successor or predecessor is repeated. Admittedly this last explanation would hardly commend itself for elementary instruction. But for any real understanding of arithmetic the grasp of an operation's being performed *so many times* is quite indispensable; and abstraction of a feature from groups of gingerbread nuts cannot give us this grasp.

Abstractionist accounts of these matters are extremely unsatisfactory. What is done in addition or subtraction can be repre-

many times some operation has been performed; having counted several times from 'one' to 'ten', we can notice how many times we have so counted, and then how many times we have counted *ten* counts from 'one' to 'ten'—and so on, and so on, *indefinitely*. This sort of pattern is a free creation of the human mind; no pattern in the natural world is there to suggest it.

In abstract counting, then, what is fundamental is the use of a numeral as (in Wittgenstein's phrase) the exponent of an operation. Counting objects of a kind adds two things to abstract counting: first, the recognition that the objects are all of one kind (as Frege puts it, all fall under the same *Begriff*); second, the setting up of a one-to-one correlation between these objects and the numerals from 'one' up to some other numeral, which is then taken to give the number of objects of that kind.

Abstractionists give a very different account of numerals. Numerals are supposed to answer to recognizable common features of sets or groups: a group may consist of similar objects, or of any old objects (as in some recent set-theoretical presentations of elementary arithmetic) without their needing to be specially similar, or again of 'abstract units' (a very mirky notion). I shall not spell out the refutation of these lines of thought: it has been given once for all in the works of Frege, particularly his *Grundlagen*.

Once we have the key idea of a numeral as the exponent of an operation, it is easy to see how the various elementary operations of arithmetic are to be regarded: multiplication is addition repeated so many times, and division is subtraction repeated so many times; exponentiation is multiplication of 1 by some number so many times over. Addition and subtraction themselves are in principle similarly explicable, in terms of how many times the procedure of passing from a number to its successor or predecessor is repeated. Admittedly this last explanation would hardly commend itself for elementary instruction. But for any real understanding of arithmetic the grasp of an operation's being performed *so many times* is quite indispensable; and abstraction of a feature from groups of gingerbread nuts cannot give us this grasp.

Abstractionist accounts of these matters are extremely unsatisfactory. What is done in addition or subtraction can be repre-

sented, in a superficially convincing way, in terms of abstracting recognizable common features from manipulations of pebbles or gingerbread nuts. But this sort of account already comes under some strain when we think of multiplication. I remember the account of money sums given in a 'set-theoretical' elementary arithmetic; it broke down completely because of course there is not, as the author's treatment required, a set that is *the* set of ten dimes answering to a given dollar bill. For exponentiation of course the case is hopeless: no manipulations of a set with 2 members and a set with 5 members will yield a set with 2^5 members.

Color Concepts

Someone might hold that abstractionism at least gives a correct account of the acquisition of *some* concepts, particularly of simple concepts tied up with sense-experience. I hold, as I did when I wrote *Mental Acts*, that abstractionism is a wholly misbegotten style of thinking and has no legitimate application at all. I imagine it will be admitted that if abstractionism does not work even e.g. for color concepts, then it is best abandoned and forgotten: so this is the case I shall now consider.

The word 'colored' is sometimes used to cover whites, blacks, and greys, sometimes to exclude them. In the latter application of the word, the adverb 'chromatically' is sometimes added for disambiguation's sake; by etymology, of course, 'chromatically colored' just means 'colorishly colored'. The use of this learned word in *Mental Acts* caused some critics to stumble strangely at my maintaining that the concept *chromatically colored* is as simple, as directly related to sense-experience, as e.g. the concept *red* or *scarlet*: surely only very sophisticated people can have *that* concept! My point is quite simple: the word 'colored' without a qualifying adverb is very often used in ordinary language in the sense I distinguished by writing 'chromatically colored': thus we speak of colored glass, colored light, colored illustrations in a book, and similarly of color prints and color television. The difference between windows with panes of plain and of colored glass, or a plain and a colored illustration, is as much a matter of direct observation as picking out something red or

scarlet; and that something is in this sense colored may be observed and remembered without our noticing particular colors. (It would thus not surprise me very much if a child should learn the word 'colored' in this sense before it learned words for particular colors; but I have no idea whether this ever happens; still less did any argument of mine about color concepts turn on this matter of child psychology, as some critics supposed.)

Now *in rebus* there certainly are not two features of a given red thing, its redness and its (chromatic) color. Being red is just the way the thing is (chromatically) colored, not something superadded to its being colored, nor yet something analyzable into color *plus* a differentia. Since there are not two features to be discriminatively attended to, there can be no performances of picking out each from the other, upon which to base our formation of the two distinct concepts *red* and (*chromatic*) *color*. But concepts of simple sensible characteristics are the abstractionist's favorite and paradigm examples; if he is wrong about these, then it is not just that he has chosen the wrong model for explaining logical or numerical or relational concepts—he is wrong from beginning to end.

The Origin of Concepts

How then are concepts got? It is not a fair challenge, in the first place, to demand of me that I give a positive account if I want to reject abstractionism. If abstractionism is incoherent, we must purge our minds of abstractionist ways of thinking. Or, to use Descartes's metaphor, if the house is about to fall down you must move out, no matter how poor a provisional shelter you move into.

Again, it might be said that the way we get concepts is a matter for empirical enquiry and philosophy cannot pre-empt the answer. We can, however, certainly require of a theory that it should give a coherent account of the performances by which we acquire an ability and of the performances in which the ability is exercised: the latter requirement is specially pressing, for only through its exercise can an ability be identified and discussed at all. Abstractionism gives no such coherent accounts, so we may dismiss out of hand any claim that there is empirical

evidence for abstractionism. People might as well claim that the bad old logic can be rehabilitated by some empirical enquiry into the way men actually think and reason (it would not surprise me if such a claim had been made).

Concepts are exercised in our rational discourse, spoken and unspoken. The human mind forms concepts, and exercises them in making judgments and statements. It is what men think or say that is true or false; concepts are abilities by which we form these pieces of discourse, just as there are distinguishable abilities by which chess positions are generated. 'What then makes pieces of discourse true or false?' We need not assume in advance that a *general* answer to this question is to be sought, any more than to the question 'What makes men happy?' On the face of it men are made happy, and judgments made true, in very various ways: it all depends. Of course philosophers have thought there *must* be a unified answer to either question. But even if there is a general answer and we do not yet know it, we are *not* going to advance towards it by going down an abstractionist *cul-de-sac*. The very same concepts may be exercised in a true bit of discourse and a false one, e.g. in 'Cats eat mice' and 'Mice eat cats'; so we need not and cannot think that men achieve a true picture of reality by the *identikit* technique of combining the recognitions of the several features involved.

The Causation of Action

G. E. M. ANSCOMBE

If we think of some commonplace happening, such as a door's shutting, we can readily imagine a range of different answers to the question "What made that door shut?" I will mention a few, and to each add some natural questioning that might follow it up.

1. An apparatus attached to the door.
 - (a) How does it work?
 - (b) Who put it there, why and when?
2. This wedge was propping the door open and got removed.
 - (a) How did it get removed?
 - (b) Why did the door shut when the wedge was taken away?
3. A blast of air blew it shut.
 - (a) Why was there a blast of air there then?
4. Its own weight caused it to shut.
 - (a) How do you know—i.e. is that certain?
 - (b) How is it hung so as to shut itself like that?
 - (c) And why was it open, then?
5. A powerful magnet pulled it shut.
 - (a) Why could the magnet affect the door?
 - (b) How did the magnet come to be there?
6. A sudden vacuum was created in the space beyond the door.

- (a) Why does a vacuum have that effect?
 - (b) What created the vacuum?
 - (c) (Given certain answers to b) Why was that done?
or How did that come about?
7. Something flew against the door and banged it shut.
- (a) What?
 - (b) What propelled this object against the door?
 - (c) Why was that enough to shut the door?
8. The dog pushed it shut.
- (a) What was the dog doing there?
 - (b) Was the dog actually trying to shut the door?
9. Jones shut it.
- (a) How?
 - (b) By accident or on purpose?
 - (c) Why?

I give this list to draw attention to the wide difference of further questions and interests naturally aroused by the different answers to the first, 'simple' question "What caused the door to shut?" The different answers as it were adjust us to a variety of new enquiries.

All the answers are perfectly appropriate. We can pick out from among them those which name causes which act on the door: the artificial mechanism, the wind, the dog, the projectile, the magnet, the human. By contrast, the removal of the wedge was a '*causa removens prohibens*'—a cause that removes a hindrance; the creation of the vacuum produced an imbalance of air pressure, as a result of which what moved the door—acted on it—was the air on the other side of it. And what would one say about the weight of the door, which caused it to shut 'of itself'? Neither that the weight moved the door nor that it did something that led to something else's moving it.

This brings out how a general enquiry into the nature of a *cause* is rather like a general enquiry into the nature of a *factor*. We may be reminded of Aristotle's four causes: he at least recognized some variety. But four is not enough. E.g. the door's weight does not belong under any of Aristotle's headings. We certainly need to remember often repeated warnings against using the expression "*the cause*." We shall prefer expressions of

the form "*p* because *q*" and "*p* because of *x*." When the magnet pulled the door shut, the door shut because of the magnet, but possibly also because the wedge was removed and certainly also because of the way the hinges were seated.

We don't ask *how* the wind blows the door shut. It shuts the door by blowing it shut. Long since, we grasped that rapidly moving air *presses*, and the citation of the wind, blowing against the door as cause of its shutting, does not evoke questions as to the mechanics of its action.

The first question about the apparatus, "How does it work?", might be echoed about the magnet. Here we want a *theory* of magnetism, a general theory of natural science. Comparably, a theory of gravity seems to dissolve our question what sort of cause weight is, whereas our question 1(a), "How does it work?", is a question what sort of apparatus *this* is—magnets and weights might come into it. Now compare "How did the apparatus work?" with "How?" asked of Jones' shutting the door. "It exerted a pushing action" would be a very inadequate answer to "How did the apparatus shut the door?", a mere slight specification of the type of action in answer to a question which is naturally a request to be told something about the inner working of the mechanism. By contrast "by pushing it" may be an adequate answer to "How did Jones shut the door?" and doesn't mean: He's a door-shutting mechanism which works by a pushing action.

Nevertheless, there is a question "How does that work?" to be asked about someone's pushing a door with his hand, say; about, that is, likely answers to g(a). —And we are *also* inclined to ask verbally the same question about likely answers to g(c): "Why did he shut the door?", e.g. "So as to have a private conversation." How does *that* work? But first let us attend to this follow-up of answers to g(a).

To repeat, we don't ask how the wind works; pressure we feel we understand as a cause of new motion. But, familiar as we are with the capacity of a human to give a push, still our curiosity is aroused how *this* mechanism may operate. The answer treats of impulses in efferent nerves, of the contraction of muscles, of the pull of muscles on tendons and bones, of ball-

and-socket joints, and so on. Now our enquiry is like that into the artificial mechanism.

There are two quite distinct directions of enquiry here. In the *first* we are interested in picking out 'chains' of causality going back in time. (*Chains* are picked out from *fans*.)

The door moved because of the push from the arm of the mechanism; that happened because of the expansion of a spring; that, because of the previous compression of the spring; that, because of the previous movement of the door in the other direction; that, because of the push of a hand; that, because of the placing of the hand and the extension of the arm; that, because of the contraction of the muscles; this last, because of the message down the efferent nerves; and that because of ____ what? No one knows in this line of causes unless he is helped by information of a different sort: it can be told that the man shutting the door was, say, obeying an order or had caught sight of something that made him want to shut the door. If so, we can go on where we broke off: that, because of the afferent nerve impulses leading to the sensory cortex and other parts of the brain; these, because of the impact of vibrating air on the ear drum etc.

Here one wants to say: there was a gap. What came between the impulses in the efferent and the afferent nerves? Well, not much is known. But—and this makes plain our *second* direction of enquiry—is there not also a gap between, say, the impulses in the efferent nerves and the contraction of the muscles? This gap, as it happens, is easier to work at filling in. The question is: how? How does the nerve affect the muscle? Such-and-such happens—a change in the character of a calcium salt, say. How does that come about, and how does it make this bit of tissue slide over that? This 'how' about the connexion between established links is the second of the two directions of enquiry.

When we want to know how a human being—or other animal—works in doing such things as pushing doors, we are usually asking for answers to questions of this latter sort. A set of questions tracing causal chains, not back and back, but in and in: until we reach, if we ever reach, elementary links—links not themselves chains.

To know that the impulses in these afferent nerves are relevant, we have to make a judgment that the action was an immediate reaction to an external stimulus; was, e.g., obeying an order just given. That is the clue that makes us attend to these particular processes of perception here, to the impulses in those nerves. Recognizing that is recognizing a pattern of a different sort from the patterns of elementary physical causation.

An analogy to illustrate this: suppose that for some purpose or other you were plotting relations of color spots in a *pointilliste* picture. Something gives you a line of spots to trace up. This something corresponds to the push of this object, the hand, against the door, and to whatever tells you to look inside the hand and arm as, in like case, you would *not* look inside a block of wood. You pursue your line of dots up to a certain point. Then you can't pursue it any further unless you step right back and see the *figure* that is painted. Doing so gives you your clue, it tells you to connect your line with *these* other spots elsewhere within the figure. Eventually, you hope, you may be able somehow to join up the two lines. But recognizing the figure was necessary; if your new line did not hang together with that, it could not be a right line. Recognizing the figure was a different sort of observation from noting the relations of the particular spots. But the concern is simply to find where to continue tracing your line of spots. Similarly the interest *here* in recalling or noting that the animal was, say, obeying an order lies in our learning from this where to look in order to continue our line of causality back from the impulses in the efferent nerves. Knowing this is of course essential to knowing what gap we have to fill in when our interest is not so much in tracing causes back and back but in asking "How does this effect that?"

There is, however, a different sort of enquiry "How does that work?", when that question was elicited by answers to 9(c), "Why did he shut the door?"; and equally when it is elicited by answers to 8(b), "Was the dog trying to shut the door?" And accordingly a different sort of answer. How does a human—or other similar—animal work? By looking for food, by recognizing danger and responding with flight or fight, by obeying orders, by calculating how to attain various ends. But now: what sort of 'ways of working' are these?

Here someone may say: "In connection with these things we can certainly ask what goes on in the brain and nervous system; what is their state, when an animal is in one of these psychological states—looking for food, for example. We can already sometimes get an answer to this. And if here too we did have to 'step back' as you put it, and 'recognize a pattern of a different sort,' that is just a *methodological point about our present situation*. When we have the information we want, there will be no need of 'stepping back' and that especially striking gap in our causal chain, which we found in the physiological enquiry, will be filled, in some cases, presumably, with the brain states corresponding to beliefs and wants. These will be states of the system either just produced or pre-existing and such that the impulses in the afferent nerves mesh in with them in discoverable ways and the impulses in the efferent nerves are then produced. In this way we are after all *not engaged in a different sort of enquiry.*"

This is a position, or complex of positions, which I will show to be wrong.

It makes the assumption that the explanation of the coming about of actions by volition and intention is what thinkers of modern times call 'causal' explanation and that this is just one single sort of explanation. And similarly for reference to what someone believes, when this comes into explanation of his action.

Not that the existence in a man of a belief, a desire, an aim, an intention, may not be causes of various things that later come about. Indeed they may, and the effect of an intention may even be an action in execution of that intention! E.g. suppose I have a standing intention of never talking to the Press. Why, someone asks, did I refuse to see the representative of *Time* magazine?—and he is told of that long-standing resolution. "It makes her reject such approaches without thinking about the particular case." This is 'causal' because it says "It makes her . . .": It derives the action from a previous state.

The mistake is to think that the relation of *being done in execution of a certain intention, or being done intentionally*, is a causal relation between act and intention. We see this to be a mistake if we note that an intention does not have to be a distinct psychological state which exists either prior to or even

contemporaneously with the intentional action whose intention it is. E.g. someone applies some extra force to a telephone dial-piece because it is a bit jammed. Is not his application of extra force perfectly intentional?—as opposed, let us say, to a case where he moves his hand (with the finger already in one of the holes) somewhat violently as involuntary recoil from a sting. But there was, we *can* suppose, no prior formation of intention, nor is the intention a mental state that accompanies the action. *That* the action under this description, “applying a bit of force,” was intentional, comes out in his explanation, in what he says if someone asks him why. “To unjam it,” he says. So the application of the extra force was a means to the end which he mentions. But that this was so was formulated *after* the event. Was what was formulated not there before? All introspection or observation can tell him, we may suppose, is that *it seemed jammed* and then he acted. He doesn’t find e.g. that the *thought* of the need to unjam it ‘went through his head.’ But didn’t he *want* it to move? Well, what was that supposed to be like? Did he *feel* something which he could call a desire that it should move, as when he is showing someone an experiment in which something is supposed to move and he watches it with anxiety? No. This is what he was doing—dialing a number. Of course he wanted it to move! But saying so does not add a new event to the record. The teleology of conscious action is not to be explained as efficient causality by a condition, or state, of desire. Remembering that *that* was ‘what I did _____ for,’ does not have to involve remembering such a state.

Suppose you say at this point “I admit in such a case I may not recall such a *thought*, or any feeling of desire, but the action’s being ‘to unjam it’ imparted a certain atmosphere, a certain character, which I *now* recall its being suffused with, and *that* was the intention.” Was that then a separable mental experience which you want to say *caused* the action? For that was what we were arguing about: whether there being such-and-such an intention—in this case, of applying force, since that is what the intentional action is—causes the action which is described as intentional. And in this conception a cause has to be thought of as a distinct thing, which is found to have this effect: as, e.g., the evocation of solemn feelings may keep me from laughing

in some game. You didn't find that when you experience *this* atmosphere it turns out that there is an action of applying extra force to a slightly stuck telephone dial. When you introduced the atmosphere, you thought of it as something inseparable, an indefinable character intrinsic to an action when it is intentional; or rather, definable only by the description of the intentional action. But such is not a cause of the action.

At this point someone may say: "Now wait! Why all this *phenomenological* investigation? Your point of departure had nothing to do with the phenomenology of intention. Remember where we were. We were talking about tracing the physiological causal chain which leads up to the door's getting pushed. You said, and we admitted, that we have to 'step back' and 'recognise a pattern of a different sort'—i.e. see that the human is acting in obedience to an order—before we know where to look to continue the causal chain back beyond the impulses in the efferent nerves. When we know that, we can find the impulses in the afferent nerves produced by the impact of sound waves on the ear-drum. We granted that this was true, but said it was a methodological fact about our present situation. When we've got the information we want, we shan't need to step back. The gap will be filled with the *brain-states* corresponding to beliefs and wants. We know that it used to be suggested that there will always be a gap, however close we get to closing it, in this chain of physical causality, and that this gap is filled by a mental, spiritual event of choice, will, intention, which determines the further physical processes, the messages down the efferent nerves. We have not stopped smiling at this naïveté. Your argument is surely directed against just such a view (long since outmoded) and has nothing to do with us."

On the contrary: it has plenty to do with that position. For that position assumes that that gap is to be filled with brain-states, or something meshing in with pre-existing brain-states, that *correspond* to beliefs and wants (some would say, *are* beliefs and wants, but that controversy does not interest us). Now there's no reason to say that, unless one is convinced that the explanation of the coming about of action by volition and intention is (a) true, (b) 'causal' in character.

What is meant by saying that explanation of actions by

intention is *true*? Not, of course, that every explanation of an action by an intention that may be offered is true—clearly there are both lies and mistakes. But that there is such a thing as intentional action and when there is, the intention or intentions involved in it *belong in an account* of such action. But the explanation is rather in terms of the future than the past. It would be too absurd to reject the notion of intention altogether as having a place in our account of action—to say that “He applied extra force to the dial on purpose, to free it, because it was stuck” is not a possible true account and partial explanation of an occurrence. At this point someone says that that is true of course, but the ‘ordinary language’ in which it is couched embodies a ‘mentalist theory’ which we want to reject. And *that* is why we want (roughly speaking) to put those brain-states into the gap.

But if it was a complete mistake to think of those explanations and accounts as ‘causal,’ then there is no reason to think of ‘the gap’ as being filled with something corresponding to or correlated with beliefs and desires. Fill ‘the gap’ if you can: that means, see if you can find a continuation of the chain of causality that bridges the gap. It is no more than another of those investigations that I called the second direction of causal enquiry—how do impulses in the afferent nerves in these circumstances produce impulses in the efferent nerves? It is no doubt a difficult and intractable problem, but it is mere naïveté after all to think that it must be filled by brain-states corresponding to beliefs and desires.

For there can be no such brain-states except in the particular case. I mean: there can be no such kind of brain-state as *the* kind of brain-state corresponding to *such-and-such* a belief in the sense of being a sufficient condition of it. “Why not? There are untold millions of possible states of the brain. So it may well be that among these is a set of states which all are, and which are the totality of, states of a brain whose owner for example believes that such-and-such.” But even on that supposition the brain-state is still not a *sufficient condition for the belief*.—Why not?—Because the belief might, be, say, a belief about banks, and a human whose brain might get into that state might never have heard of a bank. “Well,” it may be replied, “the brains of

such people never do get into any of these states. The causal conditions for getting into them exist in nature only where there are banks etc." But let us suppose a way of producing one of these states artificially, i.e. outside the circumstances in which the causal conditions occur 'naturally.' And now, consider the inference that if such a state has been so produced the subject is then in a state of belief that, say, 'such-and-such a bank in —cester is open at 5.00 P.M. on Thursdays,' though neither —cester nor banks nor clocks nor days of the week ever came into his life before, nor did he ever hear of them. The absurdity of the inference brings it out that even on the initial supposition—which there is no evidence for anyway—the brain-state is not a sufficient condition for the belief.

Nor is any other state of the person. Here we may be tempted to revert to the discarded position: "No other *physical* state, perhaps, but why not a *mental* state?" But in reply we can repeat the argument. We take it that a state is supposed to be something holding of its subject here and now, or over a period of time, without reference to anything outside that of which it holds or the time at which it holds: in particular, without reference to the history of the thing whose state it is. If that is how we understand a state, we can suppose the same state of an object in quite different circumstances and with a completely different history. If the argument does not apply to mental states that must be because they are not 'states' in this same sense. But however we decide about that, we cannot ascribe a belief like that about the bank's opening hours, to someone not living in a world of banks and clocks. Indeed we are implicitly looking away from the individual and into his world if we ascribe any belief to him. This we don't have to do for the ascription of a brain-state.

The same point holds for wants, aims; the same argument would have gone through if for a belief about the bank we had put wanting to rob the bank or be president of the bank. And the same goes for intentions, decisions, and thoughts—I mean thoughts in the sense of one's thinking something to oneself, say, on an occasion.

So: fill up that gap how you will. I mean, of course, *suppose* it filled up how you will. No way of filling it up, whether with

brain-states or (the fanciful) supposed correlates of expressions of Cartesian *cogitationes*, will fill it up with intentions, beliefs, wants, aims, volitions, or desires. For you are in pursuit of a type of causal history in which those things do not belong *at all*. I am not saying such a causal history is impossible—good luck to you in your pursuit of it if it is a serious scientific pursuit. Nor am I saying that it must be inadequate, 'leave something out,' always have 'the gap.' The kind of enquiry that looks for it might be completed. There might come a point at which there is no further puzzle about *how* each link in the causal chain produces the next one. And still nothing has been said about intentions, beliefs, thoughts or decisions.

But wait—did I not grant that the existence in a man of a belief, a desire, an aim, an intention, may be a cause of something that happens later, of actions of his for example? Certainly I did. Henry VIII longed for a son; the death of many children made him believe he had sinned in marrying Queen Catherine; he formed the intention of marrying Anne Boleyn. All this led to, helped to produce, the Act of Supremacy, to his decision to break with Rome. This is a causal history. It is merely a causal history of a different type from the physiological one. And indeed the physiological one touches it, but only at certain points. Henry signed something, let us say, and this was an episode in the above history. Ink got on a page in a certain pattern. It was deposited on the paper by a pen pushed by the royal hand. At the other end of the chain perhaps there was a noise—a courtier saying "Here, Sire" and messages up the afferent nerves. Etc. The causal histories of the two types aren't rival accounts.

But to repeat: it is one thing to say that a distinct and identifiable state of a human being, namely his having a certain intention, *may* cause various things to happen, even including the doing of what the intention was an intention to do; and quite another to say that *for* an action to be done in fulfillment of a certain intention (which existed *before* the action) is *eo ipso* for it to be caused by that prior intention.

It may seem that the case is analogous to that of an action's being done in obedience to an order. The order's being given is then a cause of the action—who would deny it? But here one

points to the order's *having* been given, in explanation of how it was brought about that this was done. Now if one can justly point to the prior existence of the intention as an *influencing* condition, in an account of how the action was brought about, then it can indeed be called a cause (as in the case of the refusal to see the *Time* magazine interviewer). The mere fact of priority is insufficient. One is tempted to think it sufficient because one does appeal to the *intention* to explain the action, and in the supposed case the intention existed before. But explanation by *intention* does not get a new character just because the intention existed before. It is just the same as when the intention is, so to speak, embodied in the action, and is not ever or only afterwards distinctly thought of. Contrast the case where, to my own irritation, I do something which I now don't mean to do—because I *had meant* to do it! Here the 'explanation by intention' is indeed a causal explanation. But it is not really an *explanation by intention* at all as we usually understand that expression—for the action is *unintentional*.

It is important here that the physical investigation of action takes as its object, i.e. as the system whose workings are being investigated, the individual human being. If, somewhat fancifully, a whole society were the unit system of physical investigation, then "A believes that the bank is open on Thursday at 5 P.M." might be supposed translatable into a statement, relating to any among a lot of possible histories of configurations of atoms whose totality added up to the existence of a society of people with A among them, with banks and with education in 'knowledge' of them.—But the unit of physical investigation of the action of a human being is the unit of physiological investigation—the individual human. And whether what a human being is doing is, say, signing a check, a petition, or a death warrant is not to be revealed by a physical investigation of what goes on within him: such descriptions are not a physiologist's concern, and so neither can he deal with the intentions implied by them.

When we consider 'the causation of action' we need to decide which sort of enquiry we are engaged in. Is it the physiological investigation of voluntary movement? I.e. do we want to know how the human mechanism works when, at a signal, the

hand pushes a pen, or perhaps a door shut? It is an enormously interesting enquiry. But that will not be our enquiry into the causation of action where our interests are in the following sort of question: What led to Jones' shutting the door then? We ascertain that he shut the door in order to have a private conversation with N. What history of actions, i.e. dealings of Jones and N with each other and with other people, of beliefs and wishes and decisions, led up to this action of shutting the door? That might be another interesting enquiry, an historical one to which knowledge of the detailed results of the first one is hardly ever pertinent.

If we now think in terms of, say, some sort of elementary particles and the operation of the fundamental forces recognized by physics, the very descriptions which occur in physiology may seem to be descriptions of shadows. I mean that the movement of a shadow has not any reality that has been left out once you have described the successive occlusion of light from a continuum of areas of a surface. Now what are we to think of the causal histories of human dealings of such a kind as we have mentioned? Are they so to speak shadows on shadows?

It may indeed be a handy way of speaking to say that a shadow with such and such a shape moved across the surface. Perhaps a more forcible analogue is that of a wave. No one attributes causal efficacy to a shadow. In a Disney cartoon, again, perhaps the mouse shatters a tea-pot with a hammer: but the reality is the successive complicated states of illumination of the screen, and we have, not efficacy, but a *picture* of efficacy on the part of a mouse with a hammer. But we really do find it scientifically convenient to speak of the causal efficacy of waves; *they not only move but 'interfere' with each other. All the same,* everyone will admit that this is just a convenient manner of speaking: if they are water waves, the description of the water masses bobbing up and down will be equivalent to the description of the waves, and the causal efficacy belongs rather to the masses of water particles in these up-and-down motions.

My question, then, is: are we to consider the causality of action, when we are talking about histories of human dealings, as just a highly convenient, nay indispensable, *façon de parler*,

such as we use also when we speak of waves as interfering with one another?

I shall call descriptions in terms which in this way merely amount to a convenient *façon de parler*: *supervenient descriptions*.

Let us first consider the supposition of strict and complete determinism in relation to the particles of which we and all things are composed, and the forces acting between them. I do not equate being caused with being determined. For a result to be determined is for no other result to have been antecedently possible. (A result might thus *become* determined at a point in time before it occurs, but not have been yet determined before that point.) Therefore in speaking of actions as being caused, brought about, by antecedent factors in the human history, I am not already settling the question of their being pre-determined. But given a strict and total determinism relative to the particles of which all things are composed, I think two things follow: one, that these descriptions of action and their causation in human (and animal) histories are *supervenient descriptions*, and two, that actions are 'determined,' in the sense that I have explained for that term.

The physicist David Bohm (in *Causality and Chance in Modern Physics* [Routledge & Kegan Paul: London, 1957], ch. 2, sec. 14) interestingly characterized what he called "the philosophy of mechanism" in such a way that 'mechanism' might be deterministic or indeterministic. In my own language, his characterization is this: there is some basic level of physical description such that all 'higher-level' descriptions are *supervenient*. Let the basic level be that of particles and fundamental physical forces. Then the forms of substances and animals and all sorts of actions and happenings will, as he puts it, be comparable to shadows.

We take Bohm's point and see that 'the philosophy of mechanism'—there is undoubtedly such a thing—may be either *deterministic or indeterministic*. On the other hand the position is not symmetrical. If you are a determinist at any level, it appears to me that you *must* be a 'mechanist' in relation to 'higher level' descriptions: you must regard them as super-

venient. Thus if you are a determinist about particles and forces you must regard as supervenient the descriptions of the actions and reactions of chemical substances, and of the actions of humans and other animals. And you must also regard them—the actions and reactions that in fact take place—as in truth determined from any previous point of time.

If, however, you are indeterminist at any level, you may or may not be a mechanist in relation to higher level descriptions. Thus determinism settles the question of mechanism, indeterminism leaves it open.

It is perfectly possible to be an indeterminist about some kind of particles (called 'fundamental') and their forces, while not being a mechanist in relation to certain higher-level descriptions; but nevertheless to be a *determinist* in relation to them. Thus one might think that the descriptions of chemical or animal forms were not merely supervenient. But that the existence and actions of all chemicals and animals that ever exist was determined—i.e. causally *necessitated*—from any previous point of time.

Something like this position, indeed, has perhaps been adopted as a result of the triumph of indeterministic physics. It manifests what I believe to be an itch for determinism which exists in the human mind. It is entertaining to read the last chapter of Richard von Mises's book, *Probability, Statistics and Truth*, and see what the intellectual position was fifty years ago. Granted that, macroscopically, determinism did not seem to be true, he tells us, still people had felt confident that *fundamentally* it was so; at the microscopic or submicroscopic level it would turn out to be true, and the macroscopic appearances would be revealed as illusory. But behold! at the submicroscopic level this turned out to be false. Thus von Mises. But what has happened since? At the submicroscopic level, people will say, to be sure physics has revealed a basic indeterminism. But no matter—macroscopically determinism can hold. For the macroscopic is the overall result of processes which are statistically constant. The *statistical* laws can't be infringed.

This little bit of thought-history does surely reveal a deterministic itch. What has got lost is the recognition, mentioned by von Mises, that in various ways the macroscopic appearances

are not of things being deterministic. However, it must be acknowledged that the position sketched is *possible*: at the macroscopic level determinism *may* hold in some immensely complicated fashion. But *why should one believe it does?* Here the itch calls to its aid a fallacious argument. Namely: the statistics are constant—therefore, determinism at the macroscopic level *must* be true; otherwise the statistics would be infringed by the operation of new, higher-level, *indeterministic* causes. Now this does not follow at all. The statistical laws would not have to be infringed by the operation of new higher level causes, even if these were indeterministic.

I mean: we just do not know whether, for example, the course taken by an animal is predetermined any time it runs about some area where the causal factors are constant. The appearance is *otherwise*; but that may be *illusory*—we ought to admit that *we do not know*. There can be no argument from the necessary preservation of statistical laws governing any submicroscopic processes that might be involved. The supposition of a large variety of possibilities is perfectly compatible with that. And the supposition of causal factors of a 'new' sort (a psychological sort, let us say), sometimes predetermining which of these possibilities is actualized, is equally compatible with it. It would only be incompatible if the same train of submicroscopic events were necessary for the same animal motion, described perhaps as one of attraction towards an object. Then indeed the supposition of 'new' causes would seem incompatible with the supposed statistical laws, as the 'new' causes would keep reflecting just *these* out of the previously random outcomes.

I hope it is now clear that there is *no need to regard* the causal histories of human dealings as supervenient descriptions: we *must* do this only on the basis of a radical physical determinism. Nothing is settled by saying this, however, as to the possibility of holding deterministic views in relation to 'human' causality. That question is left entirely open.

I have not at all dealt with something I have briefly indicated—the *explanation of action* by intention. This topic is indeed important, but it is big, and there would hardly be room to develop it here. In any case, I have indicated that it does not properly come under my title, "the causation of ac-

tion"—at any rate as moderns, rather than Aristotelians, understand the term "causation."

Let us end by considering the causalities especially involved in a history of people's dealings with one another. When such dealings concern or constitute great events, important in the history of nations, they are the greater part of what is called "History," where this is treated as the name of a subject of traditional lore and of academic study, a special discipline. But public or private, great events or small, the causalities concerned in them are much the same type. The first thing to note is: these causalities are mostly to be understood derivatively. The derivation is from the understanding of action as intentional, calculated, voluntary, impulsive, involuntary, reluctant, concessive, passionate etc. The first thing we know, upon the whole is what proceedings are parleys, agreements, quarrels, struggles, embassies, wars, pressures, pursuits of given ends, routines, institutional practices of all sorts. That is to say: in our descriptions of their histories, we apply such conceptions of what people are engaged in. In the context of such application, then, the causalities to which we ascribe such events can, so to speak, get a foothold. Given the idea of an engagement to marry, say, you can look for its causal antecedents. Or again this man was travelling from Aix to Ghent. What for? He was a messenger taking news. So in the situation in which the news was generated, and in which there was a requirement that he should take it, together with the instructions of whoever sent him off, and the exigencies of route or difficulties posed by his means of carrying out the purpose, together with accidental encounters and concatenations of events with aspects of temperament and facts of people's excitements—all these will contribute causalities of various kinds to the event of his arrival or non-arrival at his destination. The causalities will for example include negations. Because this man did not know this language, he went this way rather than that: a very different sort of causality from that of the issuing of a certain order to him at a certain moment. Or again: *because* he was quick tempered, he got into a rage *because* of a supposed insult, and *because* of that all unawares escaped certain dangers or involuntarily fell into other ones.

Mechanism and Meaning

BRUCE GOLDBERG

In recent years numerous philosophers and psychologists have contended that human beings are robots, or automata, or computing machines. Connected with this view is the further idea that to explain human behavior, or some aspect of it, is in effect to write a "program" for a machine capable of simulating that behavior. Such a program will be a characterization of the internal processes responsible for the behavior of the "device," human being or machine. As Jerry Fodor puts it:

. . . understanding the operations of a computer capable of simulating a given form of behavior is tantamount to understanding the behavior itself. ([7]:121)

The present paper offers a criticism of this view, with primary emphasis on the form in which it is defended by Fodor. He has, I believe, explored the philosophical foundations of mechanism in greater detail than many writers who share his general outlook. The paper consists of two parts. In the first I discuss Fodor's view concerning what can be learned about human beings from a machine which is able to simulate the speaking of a language. I argue that, given Fodor's idea of a machine simulation, the conclusion he draws about meaning and internal states is unjustified. In the second part of the paper I examine the theory of meaning Fodor defends, and on which his mechanistic conception of human beings rests. I try to show

that the theory is untenable, and that it is so for reasons similar to those Fodor offers against a related view of meaning.

I

The idea of "computer simulation" or "machine simulation" has, notoriously, been used in a wide variety of ways, and Fodor undertakes to clarify it. He offers the following criterion for simulation:

I propose to say that a machine successfully simulates the behavior of an organism when trained judges are unable to discriminate the behavior of the machine from the behavior of the organism in relevant test situations. ([7]:123)

Employing this criterion, Fodor criticizes Turing's test for determining when a machine can be said to simulate intelligent, or "cognitive," human behavior. [14] Turing proposed

that a machine simulation of human cognitive behavior should be considered to be successful if competent judges cannot distinguish between the machine and a person on the basis of their answers to questions of the judges' devising. ([7]:124-5)

This is unacceptable, Fodor says, since we can imagine a machine which, though it is able to answer questions put to it, cannot behave in anything like the way human beings do. That is,

Turing's test, strictly interpreted, does not provide a sufficient condition for successful simulation of human cognitive behavior [since] it could be passed by machines that are incapable of indefinitely many performances that lie well within the capacities of normal humans; performances that may well be argued to constitute types of intelligent behavior. Thus, the ability of a machine to pass Turing's test would not, by any means, entail that it could also obey such simple commands as "Mind the baby while I go shopping." This is in part because the ability of a machine to win at Turing's game would not ensure that the machine is able to integrate the results of its putative mentations with its ongoing behavior in anything like the normal human fashion. ([7]:125)

The trouble with Turing's proposal, in other words, is that it does not require that the machine be able to act in a way

consistent with its utterances. And if we imagine a radical enough disparity between what a machine says and what it does then it is clear that we can imagine a "question-answering machine" which would not be said to be capable of intelligent behavior. Thus Turing's criterion does not, though it should, rule out a machine which, though it can answer questions about how to boil water, "routinely puts the kettle in the ice-box when told to brew the tea." ([7]:127) Of such a machine, Fodor holds, it could not be said "that it can think, that it is rational, or that its behavior is intelligent." ([7]:126) That is, Turing's test does not require that the machine be able to integrate its alleged thought "with its ongoing behavior in anything like the normal human fashion," ([7]:125) and that is a condition for the simulation of intelligent human behavior.

Fodor thus concludes that Turing's question-answering machine is a very limited device, and that it is

in fact, precisely what it seems to be: a question-answering machine, a device that simulates one of the indefinitely many behavioral capacities which jointly constitute the rationality of a normal person—the ability to provide reasonable answers to questions that are put to him in his native language. It is a very long step from doing this to satisfying sufficient conditions upon the simulation of intelligent human behavior *tout court*, and it is perhaps not a step in any very clearly defined direction. ([7]:126)¹

At this point in his discussion Fodor (rightly, I think) requires a great deal of a machine for it to be said to be capable of simulating intelligent human behavior. For example, it should be able to carry out a command like "Mind the baby while I go shopping." Fodor calls this a "simple" command, but given what might be involved in minding a baby, it is clear that the machine Fodor has in mind is a very sophisticated sort of device. It must be capable of a wide variety of behaviors. Presumably it can do such things as preparing formula, changing diapers, getting a doctor, talking to a Jehovah's Witness who comes to the door, and so on. Fodor doesn't provide a

1. For a somewhat more extended version of this critique of Turing, from which Fodor's version is derived, see [10]: chapter 2, "The Imitation Game."

detailed list of the abilities such a machine would have to have, but it is certainly very large. In any event, the point I wish to emphasize in connection with Fodor's idea of simulation is that when he speaks here of the machine simulation of intelligent human behavior he has in mind a machine capable of behaving like a human being.

But when he comes to discussing a machine which is said to simulate the speaking of a language—an activity that is itself, presumably, a form of intelligent or "cognitive" behavior—Fodor demands much less of the machine than one might have expected. A machine simulates the speaking of a language, Fodor says, simply, when it can produce or "enumerate" all the sentences of the language. ([7]:138) Concerning such a machine, let us call it M_1 , Fodor comes to the central conclusion of his argument for mechanism, namely, that certain inferences may be made about M_1 's internal processes and that these inferences hold for human beings as well. Since M_1 can produce all the sentences of English, it can produce ambiguous sentences, identical strings of sound with different meanings. And since this is so, Fodor argues, M_1 must contain different internal states the output of which is the same "phonemic sequence." Consider, for example, the sentence "John likes old men and women." This can be understood as "John likes (old) (men and women)" or as "John likes (old men) (and) (women)." If M_1 can produce both of these sentences, Fodor says, it must contain within it two different causal states or causal sequences both of which result in the production of "John likes old men and women."

This is a far-reaching conclusion. From it, Fodor believes, we can learn something about the causal processes taking place in human brains when we speak. Since we, like M_1 , can produce ambiguous sentences there must be different brain states or different neural sequences responsible for causing the same set of sounds. Fodor states the argument concerning M_1 as follows:

Consider a device that purports to be . . . equivalent to some speaker in that it claims to be capable of enumerating precisely the set of sentences that the speaker will accept as grammatical. In order for the putative . . . equivalence to hold, it must be the

case that the behavioral repertoire of the machine includes a sequence that corresponds to "John likes *old men and women*." The present question, however, is whether it must include *two* such sequences, one corresponding to the bracketing (old men) (and) (women) and one corresponding to the bracketing (old) (men and women). ([7]:138; emphasis in text)

The answer to the question, Fodor believes, is yes. M_1 must contain two different internal sequences each terminating in the phonemic sequence "John likes old men and women." The machine, and so human beings, must be capable of going through these two different internal causal sequences. Each sequence must be contained in its program.

I want to point out that this conclusion about meaning and internal states is arrived at by considering a machine which, though it is supposed that it can produce the ambiguous sentence in question—indeed it is supposed that it can produce, or "enumerate," every sentence of the language—need not satisfy the conditions for successful simulation Fodor proposed earlier. Nothing, it should be noted, is said about the ability of M_1 to integrate its "utterances" with the rest of its behavior in a way like that of a human being, or about its ability to deceive trained judges in relevant test situations.

What if it cannot? That is, suppose that M_1 behaves like the earlier question-answering machine. Imagine that its sentence production does not tie up with the rest of its behavior in anything like the normal human way. Suppose, for example, that it says, from time to time, "I'd like to brew the tea," but completely ignores the offered kettle. When told that the baby is sleeping it routinely says, "Here I stand. I can do no other." And so on. If M_1 were to behave in this way, it would seem, whether or not it can "enumerate" every sentence of the language, what it is doing is not *speaking* a language. Paraphrasing Fodor's earlier criticism of Turing's question-answering machine, the sentence-producing machine Fodor describes is precisely what it seems to be, a sentence-producing machine, a device that simulates one of the indefinitely many capacities which jointly constitute the linguistic ability of a normal person. As Keith Gunderson writes, in a related context:

. . . if the case where the machine X-es is really the same and not just vaguely analogous to the case where the man X-es, then we should be safe in making further assumptions about the machine's general capabilities and performances. . . . In the case of certain computer outputs—a poem for example—we have hitherto understood the result in question to be such that its production required certain general skills or capacities on the part of human beings. And human beings who possessed such general skills or capacities could be safely assumed to be able to do a number of other things too. Hence if the machine truly writes poems—which would be the only sort of case where we would be justified in assuming that it was able to understand a language, reason, and reflect—then we would also be able to assume that the machine is also capable of a wide range of other activities, in which verbal, thinking, reasoning, and reflecting creatures are capable of participating. ([10]:48, 50-1)

It would seem then that—since *M*₁ does *not* simulate the speaking (or understanding) of a language—*whatever* might be true about the relationship between *M*₁'s sentences and its internal states, one could not expect much light to be shed on what is involved in the speaking and understanding of language by human beings.

But I think that the error in Fodor's treatment is a more serious one than merely that of selecting an inadequate model of simulation to illustrate his claim about meaning. For I want to suggest that the claim—where there are different meanings there must be different internal states—is not in fact based on considerations about machine simulation. It rests, rather, entirely on the philosophical theory of meaning Fodor adopts. According to this theory the meaning of a sentence *is* an internal state of the person or machine producing it. From this, of course, it will follow that if there are different meanings there must be different internal states. But this theory of meaning is untenable. It is so, I believe, for the same reason Fodor gives in rejecting a related view of meaning.

II

Fodor holds that the meaning of a sentence is an internal state of its "producer." He calls this internal state the "mes-

sage." ([6]:111) It is a mental structure which accompanies the physical sentence or "wave form," as Fodor calls it. ([6]:151)² On Fodor's view, when someone speaks, and means something by what he says, there is a "message" in his mind. This message is then "transformed" by mechanical operations taking place inside the speaker. The result of the transformation is the "wave form." This process is then reversed in the hearer. He "retrieves" the mental structure with which the speaker began.

But the idea that meaning something by a sentence involves transforming a message into a wave form, or "assigning a message to a wave form," ([6]:151) is untenable. It is so, I want to say, because the notion of a message is a confused one. The problems with it are exactly those Fodor himself raises against a theory of human thought proposed by the psychologist Jerome Bruner.

According to Bruner, there is a stage of human development, during childhood, when thinking is done by, or in, images—when thoughts *are* images. [1] Fodor holds that Bruner's theory is not a scientific hypothesis at all, even a false one, but is rather, strictly speaking, incoherent. It makes no sense to say that thoughts are images. The argument Fodor employs against Bruner, which he describes as "entirely convincing," is derived from Wittgenstein:

A picture of John with a bulging tummy corresponds to John's being fat. But it corresponds equally to John's being pregnant since, if that is the way that John *does* look when he is fat, it is also, I suppose, the way that he *would* look if he were pregnant. So, if the fact that John is fat is a reason to call a picture of John with a bulging tummy true, then the fact that John isn't pregnant is as good a reason to call a picture of John with a bulging tummy false. (A picture which corresponds to a man walking up a hill forward corresponds equally, and in the same way, to a man sliding down the hill backward; Wittgenstein, *Philosophical Investigations*, 139) For every reason that we might have for calling a picture true, there will be a corresponding reason for calling it false. That is, there is no reason for calling

2. Since Fodor advocates a form of the Psycho-Physical Identity Theory he believes that this internal structure is, in fact, in the brain. ([7]:107ff.)

it either. Pictures aren't the kind of things that can have truth-values. ([6]:181; emphasis in the text)

An image can't be a thought since any given mental image will be capable of a great variety of interpretations. The very same image can mean this or that, and be true or false, depending on the way it is *taken*. ([6]:191) Bruner has failed to focus on this fact. He supposes that the mental image refers by "resemblance." This view founders, however, since the same image can be said to resemble many different things. Fodor concludes that

there isn't much sense to be made of the notion that . . . entertaining an image is identical to thinking *that* such and such is the case. ([6]:181; emphasis in the text)

But, on Fodor's view, entertaining a "message" is identical to thinking that such and such is the case. In holding this, it seems to me, Fodor has failed to appreciate the scope of the argument employed against Bruner. For the conclusion of that argument is not that an image can't be a thought, but that no mental structure, or better, no structure at all, can play the role Bruner attributes to images. It is this role that is misconceived, not the particular candidate Bruner has selected to play the role.

An image, it is clear, can be taken or interpreted in a variety of ways. Fodor says, for example:

Suppose that what one visualizes in imaging a tiger might be anything from a full-scale tiger portrait (in the case of the ideticist) to a sort of transient stick figure (in the case of poor imagers like me). What makes my stick figure an image of a tiger is not that it looks much like one (my drawings of tigers don't look much like tigers either) but rather that it's *my* image, so I'm the one who gets to say what it's an image of. My images (and my drawings) connect with my intentions in a certain way; I *take* them as tiger-pictures for purposes of whatever task I happen to have in hand. ([6]:191; emphasis in the text)

This is the reason Bruner is mistaken in supposing that images refer by "resemblance." In order for an image to do this, one might say, the image itself would have to determine

what it resembles, and this it doesn't do. That is, in order for an image to have the property Bruner treats it as having *the image itself would have to determine the way it is to be taken*. But an image clearly does not determine the way it is to be taken since, as we have seen, the same image can be taken in a number of different ways. For this reason, if I understand him, Fodor concludes that "there isn't much sense" in the supposition that images are thoughts.

But this same point applies to messages as well. After all, aren't my messages *my* messages? Why don't I have the same freedom with regard to my messages that I have with regard to my images? In fact, in supposing that entertaining a message is having a thought Fodor is supposing that messages, unlike images, *do* determine their own interpretation. His idea of a message, that is, is that of a mental structure that *can* be interpreted in only one way. Images are ambiguous but messages, Fodor says, must be "ambiguity-free." ([6]:121)

It is, however, this idea, of a mental structure that can be interpreted in only one way, which is incoherent. This can be seen perhaps more clearly by noticing that Fodor has given no meaning to the terms he uses to describe the role messages are supposed to play. He says that a message "displays" the information *communicated by sentences*. ([6]:151) What does the word "display" mean here?

The question arises because Fodor says of images also, as he does of messages, that they display information. ([6]:191) But, as we have seen, the way in which messages display *cannot* be the way images display. For images don't display anything until they are taken a certain way. Messages, however, are thought of as displaying what they do *intrinsically*. They must be thought of, Fodor says, as displaying the meaning of a sentence

explicitly, in a way that the sentence itself fails to do. ([6]:114)

But why should messages differ in this respect from images? Messages, like images, are mental structures. How is it that they are immune from even the possibility of ambiguity? Indeed, what does it even mean to say that they are? For it seems clear that any structure, mental or physical, could be used, or taken, or interpreted in a variety of ways. There is no such thing as

an object that admits of only one possible use, an object which is *intrinsically unambiguous*. Fodor's entire view of language and mind would appear to rest on a notion which is unintelligible, in the same way that Bruner's notion that images are thoughts is unintelligible.

The confusion in this view of meaning can be seen further in Fodor's account of the constituents of messages. A message is said to be composed of elementary, not-further-analyzable units. ([6]:123) The meaning of a sentence, Fodor holds, is a structured arrangement of *atomic concepts* ([8]:496), but it hardly needs saying that the notion of an atomic concept has never been satisfactorily clarified. Moreover, it needs to be asked at this point: How could a view justifiably be considered to be scientific, not to say correct, which has at its foundation such a notorious philosophical dead end as that of an elementary meaning particle?³

In support of the claim that mechanism represents a viable approach to understanding the nature of language, Fodor cites the work of Noam Chomsky. Chomsky's theory of "transformational grammar," he believes, is a scientific, mechanistic account of language which is being fruitfully developed. In the light of what has been said, however, it can be seen that Chomsky's theory does not provide an illustration of the viability of mechanism. For the objections to Fodor's view apply equally well to Chomsky. Indeed, Chomsky advances the very same set of theses we have been considering. A language, Chomsky says,

3. Indeed, it is precisely this view of language—requiring intrinsically unambiguous structures composed of simple, unanalyzable units—against which the argument of Wittgenstein referred to by Fodor is directed. This point is discussed in detail by Norman Malcolm in [11]:152ff. Wittgenstein's insight that the *Tractatus* rests on the incoherent notion of a structure that "shows its sense" ([15]:4.022), Malcolm says, is the key to his subsequent abandonment of its view of language and mind:

In the *Tractatus* he had conceived that in order for thought and language to be possible there must be something (a picture, a proposition, a thought) that depicts a state of affairs in the world in so luminous a way that no room is left for differing interpretations. It would be something that, as it were, contained its interpretation, its application, in itself. In *The Blue Book*, the *Grammatik*, and the *Investigations*, Wittgenstein is saying that it is an illusion to think that there might be such a thing. ([11]:157; emphasis in the text)

is a combination of "sound and meaning." ([4]:17) It has "an inner and an outer aspect." ([3]:32) The outer aspect, the sound (Fodor's "wave form"), he calls the "physical signal" ([4]:29) or "phonetic form." ([3]:52) The inner aspect, the mental structure which is the meaning of the sentence, he calls the "deep structure":

It is the deep structure underlying the actual utterance, a structure that is purely mental, that conveys the semantic content of the sentence. ([3]:35)

The deep structure is produced "when the sentence is uttered." ([4]:17) It is "a mental accompaniment to the utterance." ([3]:34)

Chomsky is no clearer about the nature and function of this alleged structure than was Fodor. Thus, he says, in different places, that it is the meaning ([3]:38), that it *represents* the meaning ([4]:106), that it *expresses* the meaning ([4]:25), that it *specifies* the meaning ([4]:104), that it *determines* the meaning ([4]:104), that it *expresses the intrinsic meaning* ([4]:136) of the sentence, and, somewhat more expansively, that it "incorporates all information relevant to a single interpretation of a particular sentence." ([2]:16)⁴

What Chomsky actually should have said, from his own point of view, is not that this structure expresses the intrinsic meaning of the sentence, but that it *intrinsically* expresses the meaning of the sentence. It cannot express the intrinsic meaning of the sentence because, on Chomsky's view, sentences do not have intrinsic meaning. Sentences *acquire* meaning because they are associated with, by being "transformations" of, internal structures. It is the internal structure which has intrinsic meaning. Chomsky is right in denying intrinsic meaning to sentences, but he is wrong in affirming it of some other object. There is no such object. The idea of a structure with "intrinsic mean-

4. Sometimes Chomsky uses the term "deep structure" to refer to an intermediate product of the supposed transformational process. When this is so the final product is referred to as the "semantically interpreted deep structure." ([2]:29) At other times, as in the quoted passage, "deep structure" is the term for the final product itself. This perhaps partially accounts for some of the differences in role attributed to the deep structure.

ing," a structure which displays meaning "explicitly," is literally senseless.

Moreover, Chomsky, like Fodor, is altogether unclear about the constituents of this supposed object. His own detailed work has gone into developing another part of the theory. The nature of the internal structure, he says, is "being left unspecified pending further insights into semantic theory." ([5]:13) Chomsky says enough on this topic, though, to show the direction of his thinking, and the central idea is one we have already encountered. The deep structure is composed of "basic content elements," ([5]:58) "minimal 'meaning-bearing' elements." ([4]:138) A theory erected on such a foundation, it would seem, rests on air.⁵

Apart from the incoherence of its central concepts, one can see that the view Fodor defends fails to represent correctly many facts about language. Thus, Fodor holds that in order for someone to understand a sentence his internal language mechanism must construct the message by means of a series of operations. Among these operations is the "decomposing" of the sentence into its parts, the words, which are then replaced by the atomic meaning elements. This view of the supposed processes required for understanding reflects the idea that the meaning of a sentence, the message, is a "compositional function" of the meanings of its words, which are in turn compositional functions of atomic concepts. But, apart from the incoherence of the notions of an atomic concept and a message, when someone says something the meaning of his remark is not simply a "function" of the meanings of the words, any more than the humor in a remark is a function of the humor in the words, or the friendliness of a remark is a function of the friendliness of the words. Meaning, humor, friendliness—all these are functions of many things, the words, the way they are said, what they are said in response to, the relationship of the people involved. They are functions of the way what is said is "integrated," as Fodor puts it, into a flow of "ongoing behavior." ([7]:125)⁶

5. It is worth noting that while Fodor says that his view of mind is based on Chomsky's approach to language ([7]:ix) Chomsky says that his approach to language rests on Fodor's view of meaning. ([2]:161-2).

6. At times Fodor expresses dissatisfaction with the decompositional account of understanding. He says, at one point, that the processes required for under-

Fodor's claim about messages and understanding seems to get support from considering overly simple, and misleading, examples. If one restricts one's attention to such sentences as "John likes old men and women" or "A wise man is honest" it is easy to come to the conclusion that understanding a sentence requires a "construction" ([4]:168) or "computation" ([6]:117) of its meaning, employing the antecedently available meanings of the words. And then one is naturally inclined to suppose that there must be an internal "dictionary" ([8]:494) in which such meanings are stored, and "retrieval mechanisms" ([6]:115) to make them available. Consider, however, the following sentence:

(a) He didn't want to spend another year being Dr. No.

It is quite easy to think of many possible interpretations of this remark, many situations in which it might be used. The range narrows greatly, though, when it is heard following the sentence: "Sullivan decided to resign from the Censorship Board." Even if the notions of an atomic concept and an internal dictionary were intelligible, what plausible "entry" might there be for "Dr. No" from which the mechanism could "construct the message?" Yet many people who speak English would have no trouble at all with the remark in that context.

The extent to which meaning results from or depends on

standing may not involve decomposition into simple units, but may be carried out instead by means of "stereotypes, exemplars, images, or what have you." ([6]:153) In view of Fodor's critique of the image theory, this suggestion is a surprising one. And indeed, he says, concerning it:

The issues here are terribly difficult. . . . If your concept of a dog is, in large part, a representation of a stereotypic dog, how do you go about determining what *falls under* the concept? ([6]:153; emphasis in the text)

But, given this idea of what a concept is, it is not possible to determine what falls under the concept. As Fodor has shown, what a representation, exemplar, or image is of, depends on how it is taken, how it is used "for purposes of whatever task I happen to have in hand." ([6]:191) That is, if "I'm the one who gets to say what it's an image of" ([6]:191) I can't learn from it what falls under it, since it doesn't have a "range" until I give it one. And since this is so, Fodor's alternative suggestion concerning the supposed processes of understanding does not represent an improvement.

the embedding of language into the flow of human activity is perhaps more clearly seen in the following example. A general led a coup attempt that failed and he fled the country. The government, which had been shaky, gradually strengthened its position and, as a consequence, moderated its hostility toward the opposition. The general, through sympathizers, also made gestures of reconciliation and a somewhat friendlier climate emerged in the country. The government's original intention was to have the general shot if they caught him but now, eleven months later, at a press conference, the People's Prosecutor is asked: "What is your position with regard to General Barrios?" He replies:

- (b) I think that in the interest of tranquility this case ought to be de-dramatized.

In this situation it is clear what is being said. The general got the message. He knew that he was not going to be executed. Indeed, he knew many other things, e.g., that he wouldn't be tortured or imprisoned—that he was going to be treated leniently. When he returned home, he was questioned for a few hours and released, assured of "complete and total freedom on the condition that he refrain from any further political activity."

But if one imagines (b) being said in a quite different situation its meaning changes radically. Suppose it to be said during a discussion between the Prosecutor and one of his assistants. Suppose that it follows the assistant's observation: "It is clear that he must be caught and killed. Now, do we have a trial or is this one for the back room?" Obviously what (b) means and how it will be understood in any particular case depends crucially on the context in which it is spoken. But in developing his view of language Fodor appears to have ignored what he most emphasized at the outset, namely, the importance of seeing how what human beings say is integrated into the flow of their actions.

At this point a reply along the following lines might be made: "Certainly the context in which a remark is made is relevant to the way it will be interpreted by a hearer. Understanding a sentence involves many psychological processes, of

different kinds, and a complete theory of linguistic communication would include them all. Thus, in explaining why a hearer in a particular situation interprets a sentence as he does, a complete psychological theory would undoubtedly refer to such features of the situation as what the hearer's relevant beliefs are, what he is doing, what attitudes he has toward the speaker, and so on. The account of sentence understanding being offered is intended only as part of such a more comprehensive theory. Its focus is on one central aspect of the processes leading to understanding, namely, the hearer's transformation of the wave forms. And its goal is to characterize this 'purely linguistic' factor in communication, to discover the internal operations involved in carrying out these transformations." That is to say,

the internalized system of rules [governing the transformations] is only one of the many factors that determine how an utterance will be . . . understood in a particular situation. The linguist who is trying to . . . construct a correct grammar is studying one fundamental factor that is involved in performance, but not the only one. . . . There is no reason why one should not also study the interaction of several factors involved in complex mental acts and underlying actual performance, but such a study is not likely to proceed very far unless the separate factors are themselves fairly well understood. ([4]:27)

However, the only reason for thinking that there is such a factor involved in someone's understanding what is said to him—internal operations yielding messages—is that on this view of language to understand a sentence is to have a message in one's mind. Of course, were understanding identical with having a message (or a deep structure or a picture) in one's mind, there would be compelling reasons for assuming the existence of a program or grammar which governs the mental processes involved in the construction of messages. It would be natural to suppose further that there are processes in the brain causally responsible for those in the mind, that there is a "neural sub-structure" of the mental mechanism awaiting scientific discovery. But, if the argument of the present paper is correct, this analysis of the notion of understanding is unacceptable, for the

the embedding of language into the flow of human activity is perhaps more clearly seen in the following example. A general led a coup attempt that failed and he fled the country. The government, which had been shaky, gradually strengthened its position and, as a consequence, moderated its hostility toward the opposition. The general, through sympathizers, also made gestures of reconciliation and a somewhat friendlier climate emerged in the country. The government's original intention was to have the general shot if they caught him but now, eleven months later, at a press conference, the People's Prosecutor is asked: "What is your position with regard to General Barrios?" He replies:

- (b) I think that in the interest of tranquility this case ought to be de-dramatized.

In this situation it is clear what is being said. The general got the message. He knew that he was not going to be executed. Indeed, he knew many other things, e.g., that he wouldn't be tortured or imprisoned—that he was going to be treated leniently. When he returned home, he was questioned for a few hours and released, assured of "complete and total freedom on the condition that he refrain from any further political activity."

But if one imagines (b) being said in a quite different situation its meaning changes radically. Suppose it to be said during a discussion between the Prosecutor and one of his assistants. Suppose that it follows the assistant's observation: "It is clear that he must be caught and killed. Now, do we have a trial or is this one for the back room?" Obviously what (b) means and how it will be understood in any particular case depends crucially on the context in which it is spoken. But in developing his view of language Fodor appears to have ignored what he most emphasized at the outset, namely, the importance of seeing how what human beings say is integrated into the flow of their actions.

At this point a reply along the following lines might be made: "Certainly the context in which a remark is made is relevant to the way it will be interpreted by a hearer. Understanding a sentence involves many psychological processes, of

different kinds, and a complete theory of linguistic communication would include them all. Thus, in explaining why a hearer in a particular situation interprets a sentence as he does, a complete psychological theory would undoubtedly refer to such features of the situation as what the hearer's relevant beliefs are, what he is doing, what attitudes he has toward the speaker, and so on. The account of sentence understanding being offered is intended only as part of such a more comprehensive theory. Its focus is on one central aspect of the processes leading to understanding, namely, the hearer's transformation of the wave forms. And its goal is to characterize this 'purely linguistic' factor in communication, to discover the internal operations involved in carrying out these transformations." That is to say,

the internalized system of rules [governing the transformations] is only one of the many factors that determine how an utterance will be . . . understood in a particular situation. The linguist who is trying to . . . construct a correct grammar is studying one fundamental factor that is involved in performance, but not the only one. . . . There is no reason why one should not also study the interaction of several factors involved in complex mental acts and underlying actual performance, but such a study is not likely to proceed very far unless the separate factors are themselves fairly well understood. ([4]:27)

However, the only reason for thinking that there is such a factor involved in someone's understanding what is said to him—internal operations yielding messages—is that on this view of language to understand a sentence is to have a message in one's mind. Of course, were understanding identical with having a message (or a deep structure or a picture) in one's mind, there would be compelling reasons for assuming the existence of a program or grammar which governs the mental processes involved in the construction of messages. It would be natural to suppose further that there are processes in the brain causally responsible for those in the mind, that there is a "neural substructure" of the mental mechanism awaiting scientific discovery. But, if the argument of the present paper is correct, this analysis of the notion of understanding is unacceptable, for the

same reason that the analysis of the notion of meaning is unacceptable, namely, that the idea of a message, upon which both analyses rely, is itself an unintelligible idea.⁷

What is correct in Fodor's account, I think, is its emphasis on the treatment of meaning as a relational property of sentences. That meaning must be so treated follows from the fact that there are *ambiguous sentences*, that the same "phonemic sequence" can be taken or understood in different ways. Sentences, clearly, are not themselves intrinsically unambiguous. The error in Fodor's account, however, comes in his treatment of "disambiguation," in the supposition that for a sentence to be disambiguated, for it to have one meaning rather than another, it must be associated with a structure which is itself intrinsically unambiguous. That is, his mistake lies in supposing that the relation to be sought is that of the sentence to an *object*.

There is an alternative. It would be to see sentences as becoming disambiguated, as acquiring the meaning they have, because of, among other things, the relation in which they stand to the flow of action around them. That is, to see disambiguation as *positional*. A particular remark in a given situation will be taken a certain way. But this does not require that a structure be attached to it. The case is no different, I want to suggest, from the acquisition of such a thing as humor by a sentence. In a particular situation some sentence *s* will be found funny, while in other situations *s* will not be found funny. In this case, no more than in the case of meaning, is there a need for supposing

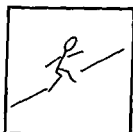
7. The parallel *Tractatus* view of understanding is described by Malcolm as follows:

To understand the meaning of a physical sentence is to come into the possession of something (*not* the sentence) the meaning of which *shows* itself—something the meaning of which is *transparent, self-revealing, unambiguous*. It is something that not only does not require interpretation but *cannot* be interpreted. It is where interpretation ends. ([11]: 140; emphasis in the text)

The notion of such a self-revealing, unambiguous structure, Malcolm shows, leads to the further idea that its constituents must be "simple" elements. ([11]: 149) I have examined some other aspects of this concept of a "meaning terminus" in [9].

that different structures attach to the sentence in the different contexts.

The idea of positional disambiguation can be illustrated by means of the example Fodor referred to earlier. This drawing is ambiguous:



1

One might see it as a man walking up or sliding down a hill. But if the picture appears in the following sequence it will not be seen in the second of these ways:



0



1



2

In this sequence it is a man walking up a hill. I don't mean, of course, that the picture can only be seen in that way. The attempt to find a picture that can only be seen in one way is, as has already been noted, a doomed one. But if this is so then the most one could get is a picture that *will* be seen in one way rather than another. And that is what we have in the above example. The picture is disambiguated because it occupies the position it does.

The moral is this: The notion of sentence meaning cannot be rendered intelligible unless sentences are seen, essentially, against the background of and embedded into the flow of characteristic patterns of human behavior. To say this is really to say nothing more than that if the life surrounding a remark were quite different from what it is, if the behavioral flow were *not* what it is, then the remark would not have the meaning it does.

This result has already been anticipated in Fodor's discussion of machine simulation. It was seen there that if a machine (or a person) did not behave in the right way, the normal human way, then it would not be speaking a language. In this sense, if there is something logically presupposed by language and meaning, it is not the existence of internal displaying structures but rather the existence and character of human life itself. For it is only given the normal flow of behavior that a sentence can have the kind of position required for it to mean something.

It may seem that a Contextual Theory of Meaning is here being contrasted with and offered in place of an Internal Structure Theory of Meaning. But that is not so. For, I want to suggest, not only is the Internal Structure Theory misconceived, but the very idea of a theory of meaning, in the sense in which Fodor tries to provide one, is itself misconceived. What is Fodor's theory supposed to do? His theory is *intended to offer a general account of how a sentence acquires the meaning it has*. It does so by association with a structure. And connected with this, as we have seen, is the idea that there is such a thing as a "complete specification," a "final analysis," of the meaning of a sentence. As against this I want to say that there is no justification for treating meaning as a property of an altogether different kind from other properties of sentences. The *remarks people make to each other are meaningful*. But they are also powerful, fanciful, witty, cautious, friendly, ironic. With respect to each of these properties it is possible to explain to someone why a remark has it. Such an explanation may be given in the form of words or gestures or pictures or some combination of these. However, and this is the essential point, there is no "preferred form" that the explanation must take. In the first place, the *kind of explanation given will depend on the kind of question being asked, on who is asking it, on what he already knows, and so on*. Explanations,

it has been observed, are "interest-relative." ([12]:41)⁸ In the second place, there is no preferred form because the idea of such a preferred form or complete specification is an unintelligible one. In this sense, there is no more a Theory of Sentence Meaning than there is a Theory of Sentence Power.

In his attempt to understand linguistic meaning Fodor, along with many other writers, seeks something at the moment of utterance, something which exists at a point-instant. This is the intrinsically unambiguous structure. If one had a "cut-out" of the mind at the moment it appears one could read off the meaning of the sentence associated with it. And one could do so, not only without knowing what the circumstances of utterance are, but even without knowing what the sentence itself is. One could say: Whatever sentence is associated with this structure means such and such. But this "static" conception of meaning fails to preserve the insight that the functioning of language, that the very idea of a language, cannot be understood unless what people say to each other is viewed as part of the continuing flow of human action and reaction. That is to say that one must look, not at a point-instant, but rather along the horizontal.

It can be seen, in Fodor's defense of mechanism, how a number of widely held views about language come together:

- (1) Sentences acquire meaning by association with internal structures.
- (2) Linguistic communication is an encoding-decoding process in which these internal structures are transformed (encoded) into sounds which are then re-transformed (decoded) into internal structures.
- (3) This encoding-decoding is carried out by mechanisms in the brain.

Fodor's treatment also reveals how these views are grounded ultimately in the concept of an intrinsically unambiguous meaning structure, a structure which represents "in so luminous a way that no room is left for differing interpretations." ([11]:157) If the argument of this paper is correct, these ideas are, strictly

8. Putnam also contends, as we have here, though for somewhat different reasons, that "'meanings' just ain't in the head" ([13]:227; emphasis in the text)

speaking, unintelligible. I believe that this is part of what it means to say that conceiving of human beings as mechanisms is a mistake.⁹

References

- [1] J. Bruner, "On Cognitive Growth," *Studies in Cognitive Growth* (New York: John Wiley & Sons, Inc., 1966).
- [2] N. Chomsky, *Aspects of the Theory of Syntax* (Cambridge, Mass.: M.I.T. Press, 1970).
- [3] ———, *Cartesian Linguistics* (New York: Harper and Row, 1966).
- [4] ———, *Language and Mind* (New York: Harcourt Brace Jovanovich, 1972).
- [5] ———, *Topics in the Theory of Generative Grammar* (The Hague: Mouton, 1969).
- [6] J. Fodor, *The Language of Thought* (New York: Thomas Y. Crowell, 1975).
- [7] ———, *Psychological Explanation* (New York: Random House, 1968).
- [8] ——— (with J. Katz), "The Structure of a Semantic Theory," *The Structure of Language* (Englewood Cliffs: Prentice-Hall, 1964).
- [9] B. Goldberg, "The Correspondence Hypothesis," *Philosophical Review* 77 (1968): pp. 438–54.
- [10] K. Gunderson, *Mentality and Machines* (Garden City, New York: Doubleday & Company, 1971).
- [11] N. Malcolm, *Memory and Mind* (Ithaca: Cornell University Press, 1977).
- [12] H. Putnam, *Meaning and the Moral Sciences* (London: Routledge & Kegan Paul, 1978).
- [13] ———, "The Meaning of 'Meaning,'" *Mind, Language, and Reality* (Cambridge: Cambridge University Press, 1975).
- [14] A. Turing, "Computing Machinery and Intelligence," *Mind* 59 (1950): pp. 433–60.
- [15] L. Wittgenstein, *Tractatus Logico-philosophicus* (London: Routledge & Kegan Paul, 1961).

9. I would like to express my gratitude to Stephen Braude for his many valuable comments and suggestions.

The Objective Self

THOMAS NAGEL

As an undergraduate I came under the memorable influence of Norman Malcolm, and the most significant thing he taught his students was probably this: that philosophical perplexity is our most important resource; that the greatest danger in philosophy is to lose the sense of what is really puzzling and so to become susceptible to answers that leave the real problems untouched. The problem I am going to discuss is one which it is very easy to 'lose' in this sense. I only hope it will stay in focus till the end of this essay.¹

I

How can I be a particular person? What kind of fact is it, if it is a fact, that I am the particular person I am? In this question the problem of putting together subjective and objective ideas about the same world takes its starkest form.

The question actually has two halves. First: how can a particular person be me? Given a complete description of the world from no particular point of view, including all the people in it, one of whom is Thomas Nagel, it seems on the one hand that something has been left out, something remains to be specified,

1. Malcolm's own views on the subject are very different from mine. See his "Whether 'I' Is a Referring Expression," in Cora Diamond and Jenny Teichman, eds., *Intention and Intentionality: Essays in Honour of G. E. M. Anscombe* (Ithaca: Cornell University Press, 1979), pp. 15-24.

namely which of them I am. But on the other hand there seems no room in such an objectively described world for such a further fact: the world as it is from no point of view seems complete in a way that excludes such additions; everything true of TN is already in it. So the first half of the question is this: how can it be true of a particular person, a particular individual, TN, who is just one of many persons in an objectively centerless world, that he is me?

The second half of the question is perhaps less familiar. It is this: how can I be *merely* a particular person? The problem here is not how it can be the case that I am this one rather than that one, but how I can be *anything as specific as a particular person in the world at all—any person*. The first question arises from the apparent completeness of a description of TN and the world which does not say whether or not he is me. This second question arises from something about the idea of 'I'. It can seem that as far as what I really *am* is concerned, any relation I may have to TN or any other objectively specified person must be accidental and arbitrary. I may occupy TN or see the world through the eyes of TN, but I can't *be* TN. I can't be a mere *person*. From this point of view it can even seem that "I am TN," insofar as it is true, is not an identity but a subject-predicate proposition. Unless you have had this thought yourself it will probably seem obscure, but I hope to make it clearer.

The two halves of the question correspond to two directions in which it can be asked: How can TN be me? How can I be TN? They are not just questions about me and TN, for any of you can ask them about himself. But I shall speak about the subject in the first person, in the Cartesian style which is intended to be understood by others as applying in the first person to themselves. It will help if you substitute yourself for me in thinking about the problem.

Let me begin with what I called the first half of the question, for its treatment will lead naturally into the second half.

II

What is the conception of the world that seems to leave no room for me? It is a familiar conception, one that people carry around

with them most of the time, of the world as simply existing, seen from no particular perspective, no privileged point of view—as simply there, and hence apprehendable from various points of view. This centerless world contains everybody, and it contains not only their bodies but their minds. So it includes TN, an individual born at a certain time to certain parents, with a specific physical and mental history, who is at this very moment thinking about metaphysics.

It includes all the individuals in the world, of every kind, and all their mental and physical properties. In fact it is the world, conceived from nowhere within it. But if it is this world, there seems to be something about it that cannot be included in such a perspectiveless conception—the fact that one of those persons, TN, is the locus of my consciousness, the point of view from which I observe and act on the world.

This seems undeniably to be a further truth, in addition to the most detailed description of TN's history, experiences, and characteristics. Yet there seems no other way of expressing it than by speaking of *me* or *my* consciousness; so it appears to be a truth that can be stated and understood only from my perspective, in the first person. And therefore it seems not to be a truth for which there is *room* in the world conceived as simply there, and centerless. If we suppose 'being me' to be any objective property whatever of the person TN, or any relation of that person to something else, we are bound to include that property or that relation in the objective conception of the world that contains TN. But then it looks as though this cannot be what 'being me' is, after all, for as soon as it has been made an aspect of the objective TN I can ask again, "Which of these persons am I?" and the answer tells me something further. Apparently no further fact expressible without the first person will do the trick: however complete we make the centerless conception of the world, the fact that I am TN will be omitted. There seems to be no room for it in such a conception.

But in that case there seems to be no room for it in the world. For when we conceive of the world as centerless we are conceiving of it as it is. Not being a solipsist, I do not believe that the point of view from which I see the world is *the* perspective of reality. Mine is only one of many points of view from which the

world is seen. The centerless conception of the world must include all the innumerable subjects of consciousness on a roughly equal footing—even if some see the world more clearly than others. So what is left out of the centerless conception—the supposed fact that I am TN—seems to be something for which there is no room in the *world*, rather than something which cannot be included in a special kind of description or conception of the world. The world *cannot* contain irreducibly first-person facts. But if that is so, the centerless conception cannot be said to leave something *out*, after all. It includes everything and everyone and what it does not include is not there to be left out. What is left out must exist, and if the world as a whole really doesn't have a particular point of view, how can one of its inhabitants have the special property of being me? I seem to have on my hands a fact about the world, or about TN, which both must exist (for how things are would be incomplete without it) and cannot exist (for how things are cannot include it).

If this problem has a solution, it must be one which brings the subjective and objective conceptions of the world into harmony. That would require an interpretation of the apparent first-person fact that TN is me and some development of the centerless conception of the world to accommodate that interpretation. If it is not a fact about the centerless world that I am TN, then something must be said about what else it is, for it certainly seems to be *true*, and from my point of view not insignificant.

There are three possible responses to this problem. One would be to solve it by discovering something about TN or about the world which is, after all, expressed by the statement that TN is me. Another would be to claim that the problem is unreal, and that the sense of a conflict between the idea of objective reality and the thought that I am TN depends on a misunderstanding of either or both of them. A third would be to find the problem insoluble, in a way that casts doubt on the idea of objective reality or the idea of the self or both.

My answer will fall into the first category. I think an account can be given of the content of the thought, which does not trivialize it. However, I would like to begin by discussing the second possibility—that the problem is in some way unreal. Of

course the most effective way of showing that it is not unreal would be to solve it, but since my confidence that I have a solution is not overwhelming I shall proceed differently. In any case, the attempts to diagnose it as a pseudo-problem are worth independent attention and may help us to understand the real problem better. I want to discuss three possible attempts, which can be called the semantic, the epistemic, and the referential response respectively. While I believe that each of them contains a good deal of truth, none of them succeeds in making the problem disappear.²

III

The semantic response is this. The first person plays an essential role in posing the problem. But there is reason to think that a misunderstanding of the logic of the first person lies behind the conviction that "I am TN" states a further fact which cannot be stated without the first person. And a consideration of how that form of words is actually used reveals that "I am TN" states no special kind of fact or truth, even though it is a special kind of statement. It is governed by truth conditions that are entirely expressible without those token reflexives.

The statement "I am TN" is true if and only if uttered by TN. The statement "Today is Tuesday" is true if and only if uttered on Tuesday. To understand the operation of such statements it is necessary only to place them in their context of utterance in an entirely centerless conception of the world; then we see that their significance and truth do not depend on the existence of special further 'facts', expressible only in the

2. Without making detailed attributions, let me mention several important articles in which the basis of such responses to the question may be found, even if the problem is not explicitly addressed. They bear on this particular topic to different degrees, but all deal with thoughts about the self: Hector-Neri Castañeda, "Indicators and Quasi-Indicators," *American Philosophical Quarterly*, 4 (1967): pp. 85-100, and "On the Logic of Attributions of Self-Knowledge to Others," *Journal of Philosophy*, 54 (1968): pp. 439-56; John Perry, "Frege on Demonstratives," *Philosophical Review*, 86 (1977): pp. 474-97, and "The Problem of the Essential Indexical," *Nous*, 13 (1979): pp. 3-21; David Lewis, "Attitudes De Dicto and De Se," *Philosophical Review*, 87 (1979): pp. 513-45; David Kaplan, *Demonstratives* (unpublished manuscript).

first person (or the present tense), which mysteriously seem to be both essential aspects of the world and completely excluded from it. The sense of these statements requires only that the world contain ordinary people, like TN, who use the first person in the ordinary way. Their sense is not the same as that of the third-person statements that express their truth conditions, but the facts that make them true or false are all expressible by such third-person statements. The rest is pragmatics.

On this view the world just is the centerless world, and it can be spoken and thought about from within partly with the help of expressions like 'I' which form the statements whose truth conditions depend on the context of utterance, a context which in turn is fully accommodated in the centerless conception of the world. Everything about the use of the first person can be analyzed without using the first person. There are no irreducibly first-person facts—only first-person statements with third-person truth conditions. The statement "I am TN" is automatically and uninterestingly true if I make it. Once we understand its logic, no further question arises as to what it says.

My objection to this diagnosis, even if it is right in denying that any further fact is involved, is that it leaves unexplained the content or apparent content of the philosophical thought that I am TN—and so doesn't make the problem go away.

There is nothing wrong with the semantic account of 'I' in itself, as one token-reflexive among others, though there is room for disagreement over the details. It tells you what you need to know about how the first person functions in ordinary communication, as when someone asks, "Who owns the blue Ford with the New Jersey license plates that's parked in my driveway?" and you say, "I do"; or when someone says, "Which of you is TN?" and I say, "I am." There is no inclination to believe that such statements express special facts: ordinary objective facts about the speaker make them true or false. Nor is the existence of any special kind of fact involved in the *making* of such statements. They are just utterances produced by ordinary individuals like TN.

But none of this either explains or exorcises the quite different thought that I have when I say to myself, looking at the world full of people saying "I own that car" or "I am his wife,"

that of all the people in this centerless world the one I am is TN: this thinking subject regards the world through the person TN. This further thought I can have even after adding to the centerless world all ordinary uses of the first person and their truth conditions—including TN's saying "I am TN," and its being true because it's TN who says it. Even after all that, there is still the further thought that TN is *me*. And when TN says to someone he meets at a cocktail party, "Hello, I'm TN," that is not the thought he is communicating. Ordinary first-person statements like "Hello, I'm TN" or "I own that car" convey information that others can express in the third person, though they are not synonymous with the corresponding third-person statements. But even when all that public information about the person TN has been included in an objective conception, the additional thought that TN is *me* seems very definitely to have further content.

On the semantic diagnosis of the problem, there is no more content for it to have: the informational content of a first-person statement can always be captured in the third person, and this one carries no such information. It is an empty use of those words, whose 'extra content' is an illusion. But as a diagnosis this does not work: no amount of concentration on the ordinary semantics of 'I' will persuade anyone who is gripped by this problem that, when he uses the word not to introduce himself but to express the philosophical thought, he is the victim of a semantic confusion and is actually saying nothing significant.

As is usually the case in philosophy, ordinary words are being used to express something beyond the bounds of ordinary discourse, but the words chosen offer themselves naturally for the expression of a philosophical thought, and are not used with a totally new meaning. This can be seen by noting that in a general way the ordinary semantics of the first person do apply here, although they do not remove the problem, because they leave the content of the philosophical thought unexplained. The thought "I am TN" is true if and only if TN has it: if I have the delusion that I am Ringo Starr, my philosophical thought as well as my self-introduction at cocktail parties will be false. But it is true that I am TN whether I am thinking it or not, and we don't yet know what makes that so. The semantic rule that "I

am TN" is true if and only if TN says it does not explain what I am saying, and it seems to me completely implausible that I am saying nothing—that it's true when I don't say it simply in the sense that it would be true if I did.

But while the semantic response does not diagnose the problem out of existence, it suggests something about what any solution should hope to achieve. It must be in some sense general. The perception that gives rise to the problem can be expressed in the first person by anyone, and not only by me. Therefore the use of 'I' here must be governed by semantic conditions general enough to be applicable to any person who can have the thought. My first understanding of it may be in application to my own case, but in some sense I also understand what someone else would mean by it. So to explain those general conditions we need an interpretation of "I am XY" which allows it to express something that each person can think truly only about himself. Any such interpretation will satisfy the condition that the thought "I am XY" is true if and only if XY has it.³

We should therefore be able to say something about the content of the first-person thought that is also comprehensible to others. We need an analogue of the informational content of ordinary first-person statements, if we are to explain why "I am TN" seems to say something about the world even when it is not just that the person speaking is called TN, or the like.

It is clear, however, that what gives the philosophical thought content must be very different from what gives ordinary first-person communications their content. It must be found by a deeper examination of how the word 'I' refers in this case, and what it refers to. Without a candidate for such an account this possibility cannot be evaluated. While the semantic response

3. Since anything that can *have* the thought is a self, we need not add that as a further condition. The Citibank cash machine that I use addresses me in the first person: "Hello, may I help you? Please put your Citicard in the slot to the right. Now tell me your I.D. code. Just a minute, I'm working on it. Sorry, I can't do that right now." Etc. It *might* say, "Hello, I'm Isabel, your friendly Citibank cash machine, at the corner of 72nd Street and Broadway." The superficial semantics of the first person apply to all these utterances. But they do not express thoughts, and the machine cannot express the philosophical thought "I am Isabel the cash machine" even if it says those words on its screen, because it is not a self.

fails to show that the problem is unreal, it remains unclear whether there can be a solution to the problem that meets these conditions.

IV

Let me now proceed to the epistemic response. The suggestion is that, although something is *expressed* by the statement "I am TN" and other first-person statements that cannot be expressed in the third person, it is not some special fact or proposition for which there is no room in a centerless conception of the world. Rather it is a special kind of knowledge or belief, namely belief about what is true of oneself, but not under a third-person description.

The only facts that correspond to such beliefs are facts about the person you happen to be, and all of them can also be stated in the third person. You might know them without knowing that the person was you—you might even possess a complete description of the world in the third person without knowing who you were in that world, and that is what gives the impression that something must have been left out by this description. But it is only a kind of *knowledge* that cannot be expressed in the third person, and this knowledge is not of some special fact or truth. It is only a special kind of attitude or mental state, a way of grasping ordinary truths, and its causes and effects are different from those of ordinary propositional knowledge. We can distinguish it from other knowledge or belief without providing it with a special object.

For instance, the knowledge that I am TN will lead me to look up when TN's name is called and to seek police protection if I overhear people plotting to murder TN. If I believe that I am about to be attacked by a bear I will curl up in a ball, whereas if I believe TN is about to be attacked by a bear without knowing I am TN I may try to warn him or get my camera. Thus beliefs about oneself play a different type of role in the determination of the individual's behavior from beliefs about the same person otherwise described. Once one has understood this, one has understood their nature, and it is clear that knowledge of who one is is not knowledge of a truth that is omitted

by a centerless conception of the world. Rather, it is a special *kind* of knowledge of perfectly ordinary truths, a mental state of a kind that anyone may have, with characteristic behavioral consequences.

But persons and their mental states and behavior, including TN's, are already included in the centerless conception of the world: that world already contains TN's knowledge that he is TN, i.e. my knowledge that I am TN, and does not require in addition the further fact that I am TN. The illusion that such a fact has been omitted from the centerless conception comes from the fact that having that centerless conception, including everything that is true of TN, does not involve being in the distinct mental state of believing I am TN.

Again I would argue that, though there is much truth in all this, it does not solve the problem. For though beliefs about oneself have special behavioral consequences, that is not all there is to them, especially not in this case. When added to a centerless conception of the world, the philosophical statement "I am TN" expresses a thought whose content is not identifiable with its tendency to produce certain behavior and expectations in conjunction with other thoughts about TN. Admittedly when I add this thought to my centerless conception of the world I am in an additional mental state; but what state is that? A causal behaviorist account of it is no more plausible than similar accounts of other mental states. Its content remains to be explained.

As was said with regard to the semantics of first-person statements, there may be ordinary cases in which the thought that I am TN or that my pants are on fire or that I am about to be attacked by a bear does not have much philosophical content. Some kinds of first-person awareness are available even to creatures without language: a cat can in some sense think "I am about to be attacked by that dog," even though it cannot think "Of all the cats in the world, I am Lucifer, the three-year-old male Persian that lives at 23 Mulberry Street." Perhaps for the cat and for the more mundane instances of human knowledge about oneself some account in terms of special attitudes may be useful. But for the philosophical thought that I am TN it will not. This is neither a special way of grasping an ordinary truth about TN—one or a number of his familiar public properties—

nor is it merely a state that shows itself in dispositions to behave in certain ways, though it *has* such consequences.

What makes the question "What kind of truth is it that I am TN?" so difficult to dislodge is that the thought that I am TN seems to be about something. Like the semantic response, the epistemic response fails to reveal this sense of content as an illusion.

V

Let me now take up the referential response, which I think was suggested to me by Keith Donnellan. According to this suggestion, the appearance that something has been left out by a completely centerless description of the world is not due to any special problem about what it is for me to be the particular person in the world that I am. We are really dealing with an instance of *something much more general*. The general point is that one can conceive of a world in universal terms without specifying any of the individuals in it—using existential quantifiers rather than names. It is possible to do this either across the board or in particular cases. I can conceive of another world qualitatively just like this one but with the identity of the individuals left unspecified.⁴ Or I can conceive of a world exactly like this one but containing entirely different individuals and therefore distinct from it. Or I can conceive of a world like this one in almost every respect, including the other individuals it contains, except that in it TN is a different individual (with exactly the same properties) from the one he is in this world. Or finally, I can conceive of a world exactly like this one except that the identity of the individual TN is left unspecified.

According to the referential diagnosis of our problem, the appearance of *something's* being left out by a centerless conception of the world is due to *something's really* being left out, by a certain form of that conception. When I detach from the fact that I am TN, I do it by imagining the world containing a TN-like individual whose identity is not specified. And relative to

4. That there is something left unspecified is the view known as *Haecceitism*. See Robert M. Adams, "Primitive Thisness and Primitive Identity," *Journal of Philosophy*, 76 (1979): pp. 5-26.

by a centerless conception of the world. Rather, it is a special *kind* of knowledge of perfectly ordinary truths, a mental state of a kind that anyone may have, with characteristic behavioral consequences.

But persons and their mental states and behavior, including TN's, are already included in the centerless conception of the world: that world already contains TN's knowledge that he is TN, i.e. my knowledge that I am TN, and does not require in addition the further fact that I am TN. The illusion that such a fact has been omitted from the centerless conception comes from the fact that having that centerless conception, including everything that is true of TN, does not involve being in the distinct mental state of believing I am TN.

Again I would argue that, though there is much truth in all this, it does not solve the problem. For though beliefs about oneself have special behavioral consequences, that is not all there is to them, especially not in this case. When added to a centerless conception of the world, the philosophical statement "I am TN" expresses a thought whose content is not identifiable with its tendency to produce certain behavior and expectations in conjunction with other thoughts about TN. Admittedly when I add this thought to my centerless conception of the world I am in an additional mental state; but what state is that? A causal behaviorist account of it is no more plausible than similar accounts of other mental states. *Its content remains to be explained.*

As was said with regard to the semantics of first-person statements, there may be ordinary cases in which the thought that I am TN or that my pants are on fire or that I am about to be attacked by a bear does not have much philosophical content. Some kinds of first-person awareness are available even to creatures without language: a cat can in some sense think "I am about to be attacked by that dog," even though it cannot think "Of all the cats in the world, I am Lucifer, the three-year-old male Persian that lives at 23 Mulberry Street." Perhaps for the cat and for the more mundane instances of human knowledge about oneself some account in terms of special attitudes may be useful. But for the philosophical thought that I am TN it will not. This is neither a special way of grasping an ordinary truth about TN—one or a number of his familiar public properties—

nor is it merely a state that shows itself in dispositions to behave in certain ways, though it *has* such consequences.

What makes the question "What kind of truth is it that I am TN?" so difficult to dislodge is that the thought that I am TN seems to be about something. Like the semantic response, the epistemic response fails to reveal this sense of content as an illusion.

V

Let me now take up the referential response, which I think was suggested to me by Keith Donnellan. According to this suggestion, the appearance that something has been left out by a completely centerless description of the world is not due to any special problem about what it is for me to be the particular person in the world that I am. We are really dealing with an instance of something much more general. The general point is that one can conceive of a world in universal terms without specifying any of the individuals in it—using existential quantifiers rather than names. It is possible to do this either across the board or in particular cases. I can conceive of another world qualitatively just like this one but with the identity of the individuals left unspecified.⁴ Or I can conceive of a world exactly like this one but containing entirely different individuals and therefore distinct from it. Or I can conceive of a world like this one in almost every respect, including the other individuals it contains, except that in it TN is a different individual (*with exactly the same properties*) from the one he is in this world. Or finally, I can conceive of a world exactly like this one except that the identity of the individual TN is left unspecified.

According to the referential diagnosis of our problem, the appearance of something's being left out by a centerless conception of the world is due to something's *really* being left out, by a certain form of that conception. When I detach from the fact that I am TN, I do it by imagining the world containing a TN-like individual whose identity is not specified. And relative to

4. That there is something left unspecified is the view known as Haeccetism. See Robert M. Adams, "Primitive Thisness and Primitive Identity," *Journal of Philosophy*, 76 (1979): pp. 5-26.

by a centerless conception of the world. Rather, it is a special *kind of knowledge of perfectly ordinary truths, a mental state of a kind that anyone may have, with characteristic behavioral consequences.*

But persons and their mental states and behavior, including TN's, are already included in the centerless conception of the world: that world already contains TN's knowledge that he is TN, i.e. my knowledge that I am TN, and does not require in addition the further fact that I am TN. The illusion that such a fact has been omitted from the centerless conception comes from the fact that having that centerless conception, including everything that is true of TN, does not involve being in the distinct mental state of believing I am TN.

Again I would argue that, though there is much truth in all this, it does not solve the problem. For though beliefs about oneself have special behavioral consequences, that is not all there is to them, especially not in this case. When added to a centerless conception of the world, the philosophical statement "I am TN" expresses a thought whose content is not identifiable with its tendency to produce certain behavior and expectations in conjunction with other thoughts about TN. Admittedly when I add this thought to my centerless conception of the world I am in an additional mental state; but what state is that? A causal behaviorist account of it is no more plausible than similar accounts of other mental states. Its content remains to be explained.

As was said with regard to the semantics of first-person statements, there may be ordinary cases in which the thought that I am TN or that my pants are on fire or that I am about to be attacked by a bear does not have much philosophical content. Some kinds of first-person awareness are available even to creatures without language: a cat can in some sense think "I am about to be attacked by that dog," even though it cannot think "Of all the cats in the world, I am Lucifer, the three-year-old male Persian that lives at 23 Mulberry Street." Perhaps for the cat and for the more mundane instances of human knowledge about oneself some account in terms of special attitudes may be useful. But for the philosophical thought that I am TN it will not. This is neither a special way of grasping an ordinary truth about TN—one or a number of his familiar public properties—

nor is it merely a state that shows itself in dispositions to behave in certain ways, though it *has* such consequences.

What makes the question "What kind of truth is it that I am TN?" so difficult to dislodge is that the thought that I am TN seems to be about something. Like the semantic response, the *epistemic response fails to reveal this sense of content as an illusion.*

V

Let me now take up the referential response, which I think was suggested to me by Keith Donnellan. According to this suggestion, the appearance that something has been left out by a completely centerless description of the world is not due to any special problem about what it is for me to be the particular person in the world that I am. We are really dealing with an instance of something much more general. The general point is that one can conceive of a world in universal terms without specifying any of the individuals in it—using existential quantifiers rather than names. It is possible to do this either across the board or in particular cases. I can conceive of another world qualitatively just like this one but with the identity of the individuals left unspecified.⁴ Or I can conceive of a world exactly like this one but containing entirely different individuals and therefore distinct from it. Or I can conceive of a world like this one in almost every respect, including the other individuals it contains, except that in it TN is a different individual (with exactly the same properties) from the one he is in this world. Or finally, I can conceive of a world exactly like this one except that the identity of the individual TN is left unspecified.

According to the referential diagnosis of our problem, the appearance of something's being left out by a centerless conception of the world is due to something's *really* being left out, by a certain form of that conception. When I detach from the fact that I am TN, I do it by *imagining* the world containing a TN-like individual whose identity is not specified. And relative to

4. That there is something left unspecified is the view known as Haecceitism. See Robert M. Adams, "Primitive Thisness and Primitive Identity," *Journal of Philosophy*, 76 (1979): pp. 5-26.

such a conception it *is* a further fact about the world that TN is the particular individual he is, namely me.

But this is not very interesting, and it certainly tells us nothing. The same is true of any individual, the eraser on my pencil for example. I can imagine a world exactly like this one in all qualitative respects, in which the identity of my eraser is left unspecified. Therefore it is a further fact about the world that my eraser is the individual that it is. But that means only that a complete conception of the world requires the specification of the individuals as well as of their properties. And this is something that can be included in a centerless conception of the world, whether the individuals are people, erasers, or anything else. The appearance that something is essentially omitted by a centerless conception is therefore due to an incomplete version of such a conception. The missing element can be supplied without going beyond the view of the world as simply *there*. And to locate the extra fact that TN is me we need only refer to his being the particular individual he is, rather than another one exactly similar.

I think there is something in this, but not yet enough to dissolve the problem. It seems to me correct that the objective conception of reality that gives rise to the problem contains an incomplete idea of TN: the problem is to say how it is incomplete.

Consider again the eraser on my pencil. It is not really such a *simple matter to conceive of a world exactly like this one* in every respect except that my pencil has a different eraser on it. The natural way to imagine the eraser different is to suppose that I took a different one out of the box last time I changed it, or that I bought a different box, or that the rubber for that batch of erasers came from a different manufacturer, or a different tree. But in all these cases I am not conceiving of the world as otherwise exactly the same *over time*, but only at the present moment. (And even at the present moment, wouldn't other things be different if a different eraser were on my pencil? What would have happened to the one that's there now?)

It seems that the only way to imagine a world exactly like this one *over time* in which there is a different eraser on my pencil is to imagine a world with a qualitatively identical history in which many of the material individuals are different. It

probably won't do even to imagine a different but qualitatively identical planet instead of the Earth, since that would have wider astronomical implications.

In other words, the identity of each material object is more or less inextricably tied in with the identity of a lot of other material objects. I bring this up in order to point out an apparent contrast with the case of the self. It is true that we can imagine a world in many respects exactly like this one except that instead of TN it contains another person exactly like him called TN, who is now writing on a topic in metaphysics. The story might be this. In 1937, my mother gave birth to two boys, developed from the same original zygote as I. One died at birth and the other was named Thomas and followed in my nonexistent footsteps to this very spot. Such an individual would not be me. Any of you who think that he *would* be me can just imagine instead that he died and the other one survived. (Actually, for all you know, it may not be I who wrote this, but rather one of those impostors.)

This is a way of conceiving that the world might contain a different individual exactly like TN by imagining the world with a different material history. But the point I wish to make is that it does not seem necessary to go to all this trouble to conceive of TN's *not* being me. Even if we imagine the world exactly as it is in all the usual respects, past and present—no twins, same ancestry going all the way back, including same individuals—even if TN is by all the usual public standards precisely the person who is now before you: still, it seems to be a quite separate further fact about him that he *is* me. That all this public history is as it has been still appears to leave something out which is, I assure you, essential to the specification of TN as he actually is. It seems conceivable that it all might have happened without my existing at all.

I say the story *appears* to leave something out. It may in fact be impossible that the objectively identifiable human being TN should have existed without being me. But if that is so, we have to explain both why it is so, and why it appears not to be. The referential diagnosis claims that my objective conception of the world leaves the identity of TN unspecified—leaves out the fact that he is the particular individual he is. But what kind of in-

dividual is that? The incompleteness seems to remain even if I do not imagine a different objective history for TN—even if I imagine the story being exactly as it was. So the omitted identity does not appear to be that of the particular physical object standing before you, or the particular biological organism, or the particular publicly identifiable person. To fill in this referential gap I need an account of what TN really is which, when I understand it, will enable me to see that that individual couldn't *not* be me. So long as it still seems like a further fact about TN that he is me, it is probable that my conception of TN is incomplete.

The natural place to look for a completion is in the identity of TN's mind. But what does that mean? I can conceive of thoughts and experiences exactly like those of TN being had by another self instead of me. It is true that the particular experiences he actually has couldn't have been had by anyone but me, if I am TN. But since I could have had entirely different experiences instead, had life gone differently, what makes TN me can't be those experiences but must instead be whatever it is that makes them mine.

So if I wish to analyze the thought that I am TN in accordance with the suggestion of the referential diagnosis—i.e. by completing the identification of the individual TN in a way that *implies* that he is me—I must find something about him more central than his body, his experiences, or even the familiar empirical capacities of his mind. I must find something that *makes* all those things mine, something that *couldn't not* be me, and that is therefore qualified to be the content of 'I' in the philosophical thought that I am TN. In other words, even if the referential diagnosis of the problem is correct in maintaining that something has been left out of the centerless conception of reality, we still don't know what that something is.

VI

I think it can be concluded that none of the three attempts to diagnose the problem out of existence is successful. The problem is real. Given all the usual things that can be said in the third person about the world, the people in it, and their experi-

ences, there appears to be something further expressed by the statement that I am TN. I am now going to try to say what this is, by asking what kind of thing the *I* is which the problematic statement asserts to be TN, or to see the world through TN—and by attempting to explain how this form of reference gives the judgment content.

To do this I must turn to what I called the second half of the question, the half that asks not how a particular person, TN, can be me, but rather how I can be anything so *specific* as a particular person at all (TN as it happens).

How can this possibly be puzzling? What else could I be but a particular person?

It is puzzling, for one thing, because my being TN (or whoever I in fact am) seems accidental, and my identity can't be accidental. So far as what I am essentially is concerned, it seems as if I just *happen* to be the publicly identifiable person TN: as if what I really am, this conscious subject, might just as well view the world from the perspective of a different person. The real me occupies TN, so to speak; or the publicly identifiable person TN contains the real me.

But since anyone who thinks TN is what I am will not be persuaded by an argument that depends on the premise I could conceivably be someone else, let me try to argue less abstractly, in an effort to evoke the operative conception of the self.

From a certain point of view my connection with TN seems arbitrary. To arrive at the problem I begin by considering the world as a whole, as if from nowhere, and in those vast spaces TN is just one person among countless others, all equally insignificant. Taking up that impersonal standpoint produces in me a sense of complete detachment from TN. How can I, who am thinking about the entire, centerless universe, be anything so specific as *this*: this measly creature existing in a tiny morsel of space and time, with a definite and by no means universal mental and physical organization? How can I be anything so small and concrete and specific?

I know this sounds like metaphysical megalomania of an unusually shameless kind. Merely being TN isn't good enough for me: I have to think of myself as the world soul in humble disguise. In mitigation I can plead only that the same thought is

dividual is that? The incompleteness seems to remain even if I do not imagine a different objective history for TN—even if I imagine the story being exactly as it was. So the omitted identity does not appear to be that of the particular physical object standing before you, or the particular biological organism, or the particular publicly identifiable person. To fill in this referential gap I need an account of what TN really is which, when I understand it, will enable me to see that that individual couldn't *not* be me. So long as it still seems like a further fact about TN that he is me, it is probable that my conception of TN is incomplete.

The natural place to look for a completion is in the identity of TN's mind. But what does that mean? I can conceive of thoughts and experiences exactly like those of TN being had by another self instead of me. It is true that the particular experiences he actually has couldn't have been had by anyone but me, if I am TN. But since I could have had entirely different experiences instead, had life gone differently, what makes TN me can't be those experiences but must instead be whatever it is that makes them mine.

So if I wish to analyze the thought that I am TN in accordance with the suggestion of the referential diagnosis—i.e. by completing the identification of the individual TN in a way that *implies* that he is me—I must find something about him more central than his body, his experiences, or even the familiar empirical capacities of his mind. I must find something that *makes* all those things mine, something that *couldn't not* be me, and that is therefore qualified to be the content of 'I' in the philosophical thought that I am TN. In other words, even if the *referential diagnosis* of the problem is correct in maintaining that something has been left out of the centerless conception of reality, we still don't know what that something is.

VI

I think it can be concluded that none of the three attempts to diagnose the problem out of existence is successful. The problem is real. Given all the usual things that can be said in the third person about the world, the people in it, and their experi-

ences, there appears to be something further expressed by the statement that I am TN. I am now going to try to say what this is, by asking what kind of thing the *I* is which the problematic statement asserts to be TN, or to see the world through TN—and by attempting to explain how this form of reference gives the judgment content.

To do this I must turn to what I called the second half of the question, the half that asks not how a particular person, TN, can be me, but rather how I can be anything so *specific* as a particular person at all (TN as it happens).

How can this possibly be puzzling? What else could I be but a particular person?

It is puzzling, for one thing, because my being TN (or whoever I in fact am) seems accidental, and my identity can't be accidental. So far as what I am essentially is concerned, it seems as if I just *happen* to be the publicly identifiable person TN: as if what I really am, this conscious subject, might just as well view the world from the perspective of a different person. The real me occupies TN, so to speak; or the publicly identifiable person TN contains the real me.

But since anyone who thinks TN is what I am will not be persuaded by an argument that depends on the premise I could conceivably be someone else, let me try to argue less abstractly, in an effort to evoke the operative conception of the self.

From a certain point of view my connection with TN seems arbitrary. To arrive at the problem I begin by considering the world as a whole, as if from nowhere, and in those vast spaces TN is just one person among countless others, all equally insignificant. Taking up that impersonal standpoint produces in me a sense of complete detachment from TN. How can I, who am thinking about the entire, centerless universe, be anything so specific as *this*: this measly creature existing in a tiny morsel of space and time, with a definite and by no means universal mental and physical organization? How can I be anything so small and concrete and specific?

I know this sounds like metaphysical megalomania of an unusually shameless kind. Merely being TN isn't good enough for me: I have to think of myself as the world soul in humble disguise. In mitigation I can plead only that the same thought is

available to any of you. You are all subjects of the centerless universe and mere human or Martian identity should seem to you arbitrary. I am not saying that I individually am *the* subject of the universe: just that I am *a* subject that can have a conception of the centerless universe in which TN is an insignificant speck, who might easily never have existed at all. The self that seems incapable of being anyone in particular is the self that apprehends the world from without rather than from a standpoint within it. But there need not be only one such self.

The picture is this. Essentially I have no particular point of view at all, but apprehend the world as centerless. *As it happens* I ordinarily view the world from a certain vantage point, using the eyes, the person, the daily life of TN as a kind of window. But the experiences and the perspective of TN with which I am directly presented are not the point of view of the true self, for the true self has no point of view and includes in its conception of the centerless world TN and his perspective among the contents of that world.

VII

Let me give a name to this subject with which I identify essentially in contrast to the apparent arbitrariness of my more familiar identity. I shall call it the *objective self*, because it is the self that apprehends objective reality. The objective self views the world through TN. I believe this picture is essentially correct, and that it is expressed by the philosophical form of the judgment that I am TN.

How do I distinguish the objective self from the person TN? By treating the individual experiences of that person as data for the construction of an objective picture. I throw TN into the world as a thing that interacts with the rest of it, and ask what the world must be like from no point of view in order to appear to him as it does from his point of view. For this purpose my special link with TN is irrelevant. Though I receive the information of his point of view directly, I try to deal with it for the purpose of constructing an objective picture just as I would if the information were coming to me indirectly. I do not

give it any privileged status by comparison with other points of view.⁵

This naturally is an idealization. Much of my conception of the world comes directly from what TN delivers to me. I have had to rely heavily on TN's experience and education, and I do not constantly subject each of his pretheoretical beliefs to detached assessment. But in a very general way, I try to do with his perspective on the world what I could do if information about it were reaching me thousands of miles away, not pumped directly into my sensorium but known from outside.

The objective self should be able to deal with experiences from any point of view. It in fact receives those of TN directly but it treats on an equal footing those it receives directly and those others it learns about only indirectly. So far as its essential nature is concerned, it could base its view of the world on a different set of experiences from those of TN, or even none at all coming directly from a perspective within the world, for in itself it has no such perspective. It is the perspectiveless subject that constructs a centerless conception of the world by casting all perspectives into that world.

Suppose all the nerves feeding sensory data to my brain were cut but I were somehow kept breathing and nourished and conscious. And suppose auditory and visual experiences could be produced in me not by sound and light but by direct stimulation of the nerves, so that I could be fed information in words and images about what was going on in the world, what other people saw and heard, and so forth. Then I would have a conception of the world without having any perspective on it. Even if I pictured it to myself I would not be viewing it from where I was.⁶ It might even be said that, in the sense in which I am now TN, I would under these circumstances not be anyone.

As things are, the objective self is only part of the point of view of an ordinary person, and its objectivity is developed to different degrees in different persons and at different stages of

5. This idea of the objective self has affinities with the 'metaphysical subject' of Wittgenstein's *Tractatus* 5.641.

6. For a similar fantasy see Daniel C. Dennett's "Where Am I?" in *Brainstorms* (Montgomery, Vt.: Bradford Books, 1978).

life and civilization. The basic step which brings it to life is not complicated and does not require advanced scientific theories: it is simply the step of conceiving the world as a place that includes the person you are within it, as just another of its contents—conceiving yourself from outside, in other words. Someone is *doing* that, and you realize that it must be you, and that you have stepped nimbly away from the unconsidered perspective of the particular person you thought you were.

Next comes the step of conceiving from outside all the points of view and experiences of that person and others of his species, and considering the world as a place in which these phenomena are produced by interaction between these beings and other things. That is the beginning of science. And again you realize that someone is doing this stepping back, not only from an individual viewpoint but from a specific type of viewpoint—and that again it must be you.

Because a centerless view of the world is one on which different persons can converge, there is a close connection between objectivity and intersubjectivity. By placing TN in a world along with everyone else I pursue a conception of him and his point of view that others may share. At the first stage the intersubjectivity is still entirely human, and the objectivity is correspondingly limited. The conception is one that only other humans can share. But if the general human perspective is then placed in the same position as part of the world, the point of view from which this is done must be far more abstract, so it requires that we find within ourselves the capacity to view the world in some sense as very different creatures also might view it when abstracting from the specifics of their type of perspective. The pursuit of objectivity requires the cultivation of a rather austere universal objective self.

It is clear from all this that the objective self that I find viewing the world through TN is not unique: each of you has one. Or perhaps I should say each of you *is* one, for I do not mean to imply that the objective self is a distinct entity. Each of us, then, in addition to being an ordinary person, is a particular objective self.

I believe this accounts for the content of the philosophical thought we have been trying to track down. It is *qua* subject of

this impersonal conception of the world that I refer to myself as 'I' in thinking the philosophical thought "I am TN." Though the 'I' is still essentially indexical, the content of the thought is that this impersonal conception of the world is attached to the perspective of TN and is developed from that perspective.

This is a solution to the original problem in the following sense. It explains how the thought "I am TN" can have a content that is nontrivial and indeed rather remarkable: almost as remarkable as it seems at first. And while it does not translate the thought into one about the world objectively conceived, it does identify an objective fact corresponding to the thought, which explains how it can have a content interesting enough to account for its philosophical 'flavor'. Because TN possesses or is an objective self, I can state a surprising identity by referring to myself indexically under that aspect as 'I', and again under the objective aspect of the publicly identifiable person TN—and I can make both references simply in virtue of possessing an objective conception of the world that contains TN. Because even an objective conception has a subject, it allows me to bring the subjective and objective views together.

VIII

This problem has much in common with other cases where informative identity statements cannot be easily explained in terms of facts about the world. What kind of fact is it, for instance, that *Hesperus* is *Phosphorus*, or that water is H_2O ? If these are identities, and their terms are not definite descriptions but rigid designators,⁷ they seem to correspond only to the "fact" that Venus is identical with itself, or that water is the substance that it is. To explain why the statements are nevertheless not trivial it is necessary to give an account of how the terms refer—an account of our different types of relation to the things we talk about which explains the significance of the statements. There are rival theories about these matters, but they all attempt to put us into an objectively comprehensible relation to the things we are talking about.

7. See Saul Kripke, *Naming and Necessity* (Cambridge, Mass.: Harvard, 1980).

The thought "I am TN" presents a similar problem, though the task is not to explain my dual relation of reference to something outside myself, but rather my dual relation to the entire world. In a sense there are two forms of reference to TN here, and we must explain the first-person reference in this philosophical context without trivializing the thought. What happens when I consider the world objectively is that an aspect of my identity comes into prominence which was previously concealed, and which produces a sense of detachment from the world. It then comes to seem amazing that I am in fact attached to it at any particular point. The content of the thought that I am TN can be understood once the objective conception closes over itself and its subject.

In a sense any purely objective conception of the world is incomplete, for it does not indicate how it is connected to the world.⁸ Since it must have a subject that is in the world, however, this automatically provides a point of contact, which is expressed in the philosophical thought "I am TN." The 'completion' essentially involves an indexical reference, though it is possible because of objective facts about TN that can be explained in the third person. Having conceived of the world impersonally, I complete my idea of TN by identifying him as the location of my impersonal conception. This explains both what it is for TN to be me and why it appears arbitrary and even accidental that I am he.

It is not, of course, accidental. In a sense, my account of "I am TN" depends on a claim about the real nature of ordinary persons. The ordinary idea of them as individuals identified by their public personae and their specific perspectives on the world is incomplete, in an essential way. Each of them, each of us, has at his core an objective self, which makes possible the reference of the philosophical 'I' and forms an essential aspect of what

8. Hans Sluga has pointed out to me that even an objective conception of the most general kind must presuppose an indexical element: for example the individuals who adopted the Gregorian calendar had to say, "It is *now* 1582"—and that is why we can describe the world in terms of Gregorian dates. The same must be true of other general terms. But the specific location of oneself seems to be something still further, not included among all these other connections.

that person is. If that is so, then I really am TN, not merely something contingently connected to him; and the apparent conceivability of my being someone other than TN is *only* apparent.

IX

I intend this to be a universal claim. But it has been put to me that my strong identification with the objective self is not universally shared—that many people regard this aspect of themselves as peripheral and inessential, by contrast with their senses, their tastes, their special qualities of feeling. The thought that 'I' in "I am XY" should refer in so impersonal and detached a way is very foreign to them, as is the sense of amazement that they should be anything so specific as a particular person at all.

It may be that I am susceptible to this identification only because the sense of detachment is an unusually prominent feature of the subjective quality of my experience. This raises the question whether alternative aspects of the self can serve just as well to explain the reference of 'I' for the purpose of the philosophical thought we have been examining. Can someone who fails to identify with the objective self even have this thought?

I think he can have part of it. After referring to himself via some subjective aspect of experience—perhaps the particular feelings or sensory experiences to which he is most attached—he can locate this 'I' as a particular person in his objective conception of the world, thus bringing the two standpoints into conjunction. But for this purpose the 'I' is not even necessary. The same thing can be accomplished by an indexical reference to something else with which the person is acquainted: "This clock is the clock in XY's office." There are many points at which subjective and objective conceptions of the world may be brought into contact.

However, that is only part of what is accomplished by the thought "I am TN." These other versions do not address the second half of the problem, which is how I can be anything as specific as TN. The objective self is unique in the following way. It is the only significant aspect under which I can refer to myself subjectively that is supplied by the objective conception of the world alone—because it is the subject of that conception.

And it is the only aspect of myself that can seem at first only accidentally connected with TN's perspective—a self that views the world *through* the perspective of TN. I believe the possibility of this self-locating thought reveals something about us all, and not only about those who find it remarkable.⁹

9. I am much indebted to Samuel Scheffler, John Searle, Thomas Sorel, David Velleman, and Eric Wefald for criticisms and suggestions.

On an Argument for Dualism

SYDNEY SHOEMAKER

I

It is a striking fact about contemporary philosophy of mind that, while scarcely anyone thinks that it is a live possibility that a mind-body dualism anything like Descartes's is true, considerable effort continues to be spent on the construction, consideration, analysis, and refutation of arguments in favor of such dualistic positions. While much of this activity takes the form of Descartes scholarship, the amount of it cannot be explained as simply a manifestation of a widespread historical interest in Descartes's philosophy; on the contrary, it seems plausible to suppose that the interest in Descartes's philosophy is due in part to the continued fascination of philosophers with the dualistic outlook of which Descartes is the classical proponent. Nor can it be supposed that the interest in dualism among philosophers is to be explained by the prevalence of dualistic convictions among non-philosophers. While there are pockets of dualistic belief in the general populace, and even among neurophysiologists, such belief seems out of tune with the intellectual temper of the times. In popular writings on scientific topics relating to the nature of man it tends to be simply taken for granted that dualism is no longer credible; and it seems a safe bet that most readers of such writings have little inclination to question this assumption. It seems unlikely, to say the least, that this intellectual climate is the result of the antidualistic efforts of such philosophers as Wittgenstein, Ryle, and Strawson. As Richard Rorty remarks in a recent paper, "The reason Cartesian dualism is so unpopular

nowadays is not because of any applications of the powerful methods of modern analytic philosophy, but simply because we keep reading in *Life* and *The Scientific American* about cerebral localization, the production of any desired emotion, thought, or sense impression by the insertion of electrodes, and the like."¹ Nor is it only *recent* developments in the biological sciences (or exaggerated popular accounts of them) that have led to the unpopularity of Cartesian dualism; as a theory about man as a natural phenomenon, dualism has been increasingly incredible since the time of Darwin.

It is worth asking, in the light of this, why it is that the attitude of philosophers towards dualism is not more like that of chemists towards phlogiston theory—i.e., why dualism is not regarded as a discredited theory whose current interest can only be historical. But while I hope that this paper will throw some light on this question, this is not its primary purpose. On the whole the paper will be an exemplification and illustration, rather than an analysis, of this *prima facie* perverse philosophical preoccupation with mind-body dualism.

I wish to consider an argument in favor of dualism which is the subject of a recent paper by Norman Malcolm.² Malcolm reports that the argument was brought to his attention by an unpublished paper by Robert Jaeger, and I shall refer to it as the "Jaeger-Malcolm argument." It should not of course be assumed that either Malcolm or Jaeger actually subscribes to the argument; their interest in it is basically of the same sort as mine, the nature of which will soon be apparent. Malcolm's own analysis and evaluation is subtle and complex, and any brief summary of it would be bound to distort it. It is different enough from my own that a consideration of it would take me away from the issues I want to consider, so I shall not attempt to discuss it here.

I also shall not consider the historical question of whether Malcolm is correct in attributing the argument to Descartes.

1. Richard Rorty, "Functionalism, Machines, and Incommensurability," *Journal of Philosophy*, 69, no. 8 (April 20, 1972): pp. 218–19.

2. Norman Malcolm, "Descartes' Proof that He Is Essentially a Non-Material Thing" in *Thought and Knowledge, Essays by Norman Malcolm* (Ithaca: Cornell University Press, 1977), pp. 58–84.

While the attribution strikes me as plausible, it does not matter for my purposes whether it is correct. Whether or not the Jaeger-Malcolm argument is Cartesian, it is certainly "Cartesian." This is really an autobiographical remark; what I mean is that the argument strikes a responsive chord in me which is akin to that struck by certain passages in Descartes's writings. While I am as certain as I am of anything that mind-body dualism has to be false, I am nevertheless aware of a suppressed inclination in myself to believe that it is true. There is, as it were, a tiny dualist faction in my soul which has not been completely quieted by the much larger materialist faction. And the Jaeger-Malcolm argument appeals to the dualist faction in me in much the way the dualistic passages in Descartes do. While such philosophical inclinations are bound to be multiply overdetermined, it strikes me as possible, or even likely, that my produalist inclinations are partly to be explained by the initial plausibility, the initial semblance of soundness, of this argument. This of course involves supposing that a person's beliefs and inclinations to believe can be influenced by arguments which he has never explicitly formulated. But I am sure that this is true. An important part of philosophical activity consists in making explicit, so that they can be subjected to analysis and criticism, lines of reasoning which have previously been only implicit in our thinking. In any case, I am sure that I am not alone among contemporary antidualist philosophers in having some produalist inclinations. Many of us have within us a little dualist which we would like to exorcise; and the way in which we hope to do this is by bringing to light the sources of our produalist inclinations, so that these can be exposed as involving confusions, misconceptions, fallacious reasoning, and the like. As I have indicated, it would be unrealistic to hope that one might discover some *single* argument or line of reasoning which is the sole source of one's produalist inclinations and is such that once one has seen it refuted the inclinations will wither away. It is more likely that these inclinations have a plurality of causes, some of which, perhaps, have little to do with reason and argument. Nevertheless, the Jaeger-Malcolm argument strikes me as a likely candidate for being one of the sources of produalist inclinations, including my own, and to warrant our attention for that reason.

It is worth remarking, before we proceed to an examination of the argument, that the Jaeger-Malcolm argument resembles other arguments in favor of dualism which have commanded the attention of philosophers in being, at least *prima facie*, of an *a priori* character. It has, in fact, been quite common in recent philosophy to treat the issue of whether dualism is true as one to be settled by *a priori* reasoning. To be sure, dualism is occasionally argued for on empirical grounds. Spiritualistic phenomena might, if taken to be genuine, be regarded as supporting dualism. And occasionally a neurophysiologist attempts to argue that the mechanisms in the brain are not up to the task of doing all that the mind does, and that a nonphysical mind must therefore be postulated.³ But it is not such considerations that motivate the little dualist in me, and I suspect that the same holds for most other philosophers interested in dualism. Insofar as we are tempted to believe in dualism, we are so tempted on the basis of considerations that are *a priori* rather than empirical. Nor is it only *pro*dualist arguments that tend to have an *a priori* character. The *a priori* reasonings of *anti*dualist philosophers have often been directed at showing on *a priori*, or "conceptual," grounds that dualism is false—or senseless, or conceptually incoherent—and not merely that the *a priori* arguments offered in its favor are fallacious or confused. This is rather striking in the light of the fact, mentioned earlier, that the *anti*dualism which characterizes the present-day intellectual climate seems to be due mainly to such scientific developments as Darwinian evolutionary theory, the discovery of the physical basis of heredity, the vast increase in knowledge of the workings of the brain, and advances in the field of artificial intelligence and computer simulation of human intellectual functions. Surely philosophers have been influenced by these developments at least as much as anyone else. But it seems odd that someone should be strongly inclined to reject a view on empirical grounds (e.g., because of discoveries in biology) while also being inclined to accept that same view on *a priori* grounds. And it seems even

3. See, for example, Wilder Penfield, *The Mystery of the Mind* (Princeton: Princeton University Press, 1975), and John Eccles's essay in Karl R. Popper and John C. Eccles, *The Self and Its Brain: An Argument for Interactionism* (Berlin: Springer Verlag, 1977).

odder that someone who rejects a view on empirical grounds should feel that it ought to be possible to give an a priori disproof of that view. This gives rise to a doubt as to whether the dualism which philosophers talk about can really be the same as the dualism which philosophers and educated laymen alike find unbelievable because of developments in empirical science. And perhaps this helps to explain the preoccupation with dualism that I mentioned at the beginning of this paper. The produalist inclinations that some philosophers find in themselves are not merely an intellectual itch which those philosophers would like to be rid of; the existence of these inclinations, and the fact that they go with an inclination to regard the truth or falsity of dualism as something to be settled by a priori reasoning, e.g., by "conceptual analysis," are signs that we don't really have a very clear idea of what we are denying when we deny that dualism is true. But if our idea of this is unclear, so is our idea of what we are affirming when we assert that materialism, or physicalism, is true; for as far as intelligibility is concerned, these doctrines take in one another's washing.

II

I turn now to a consideration of the Jaeger-Malcolm argument. It is formulated by Malcolm as follows:

- (1) *I think I am breathing* entails *I exist*
- (2) *I think I am breathing* does not entail *I have a body*

Therefore

- (3) *I exist* does not entail *I have a body*.

The first thing to be said about this argument is that it is logically valid; it is impossible that its premises should be true without its conclusion being true. It is an instance of the formally valid argument form:

p entails *q*

p does not entail *r*

Therefore, *q* does not entail *r*.

So any successful challenge to the argument will have to be a challenge to one of its premises.

But before we begin to consider such challenges, we should notice that on the face of it the argument falls short, even if sound, of establishing the Cartesian conclusion that one's mind, or self, is something distinct from any body—that it is a spiritual, incorporeal substance lacking all physical characteristics. Proposition (3), the stated conclusion of the argument, is logically equivalent to the proposition "It is logically possible for me to exist without having a body." This does not of course entail that I *do* exist without a body, and Descartes would not suppose that it does; it was of course no part of his purpose to deny that he *had* a body. But given that I have a body, why shouldn't it be the case that right now, while I have it, I possess its corporeal characteristics, even if, as (3) says, it is possible that I should exist without having such a body, and so without having any physical characteristics? In other words, one might suppose that just as the claim that I could exist without having a body is compatible with my actually having one, so the claim that I could exist without having physical characteristics is compatible with my actually having such characteristics. But if the latter is so, then if we take dualism to imply that I, as the subject of my mental states, am something nonphysical or noncorporeal, something not having any physical characteristics, then the conclusion of the Jaeger-Malcolm argument does not imply dualism.

But there are strong *prima facie* objections to such an attempt to reconcile acceptance of (3) with the rejection of dualism. If right now, while I have a body, I necessarily have the physical characteristics which that body has, it would seem that I ought to be identical to that body. But how could it be true both that I am identical to that body and that it is possible for me to exist without it existing? It can be replied that I needn't be identical to my body in order to have its physical characteristics. Perhaps my relationship to my body is like that of a ship to the particular collection of wooden planks that constitute it at a given time, or of a statue to the portion of bronze that constitutes it. Since the ship can survive the replacement of some of the planks, or even all of them if the replacement is gradual enough, it is not identical to the collection of planks; but its shape, size, etc. at a given time is precisely that of the collection

of planks that constitute it at that time. In an analogous way, it might be suggested, I might share my body's corporeal characteristics and yet not be identical to it. I think that this suggestion is correct. But this analogy does not help us make sense of the claim that I might at one time have all of the physical characteristics of a certain body and at a later time have no physical characteristics at all—for if we substitute "collection of planks" for "body," this is plainly not something our ship can do.

There is, indeed, one way in which we can make sense of this claim—but it is one that is of no help to someone who is out to reconcile (3) with the rejection of dualism. Descartes himself can perfectly well allow that there is a sense in which I am the subject of whatever corporeal characteristics my body has, even though he thinks I can exist without having any body at all. Descartes can agree that if my body weighs 170 pounds, I weigh 170 pounds. But his view must be that I have these characteristics in a derivative way. They do not belong to myself, that which I call "I," in the direct way in which my mental characteristics belong to it. What they directly belong to is the body with which I am "intimately united," and what directly belongs to me is the relational property of being united to a body which has such characteristics. A Cartesian can hold that we sometimes use predicates like "weighs 170 pounds" to ascribe such relational properties; and when such a predication is true, the subject may be said to have the nonrelational property in question—in this case, weighing 170 pounds—"derivatively." To have a property derivatively is to have the relational property of being related in a certain way—roughly, the way Descartes thinks souls are related to their bodies when they are embodied—to some other entity which directly (and so nonderivatively) has the property.⁴ Descartes's dualist doctrine is that he, what he calls "I," is something that does not directly, or *nonderivatively*, have any corporeal characteristics. And while it is *prima facie* intelligible that something should at one time have physical charac-

4. I would not want to say that the ship possesses only derivatively the properties that belong to the collection or aggregation of planks that constitute it at a particular time; the relation of the ship to the collection of planks is obviously of a very different sort from the relation Descartes believes to hold between a person and his body.

teristics derivatively and then subsequently cease to have any—at any rate, we have a picture, of the soul being released from the body, which goes with this—it is much harder to understand the suggestion that something could at one time have physical characteristics in a *nonderivative* way and then cease to have any.

It is clear enough, in any case, what we have to add to our argument in order to get the explicitly dualistic conclusion which Descartes could express by saying "I am something incorporeal—something having no corporeal, or physical, characteristics." What we have to add is what I shall call the "essentialist premise." This says that whatever has corporeal characteristics *nonderderivatively* is essentially, or necessarily, something that has corporeal characteristics. In other words, if something has, *nonderderivatively*, any corporeal characteristics, then it is not possible for it to exist without having some corporeal characteristics or other. This is, it should be observed, a principle which Descartes was committed to in any case, and certainly would have had no hesitation about using. Spatial extension being, on his view, the essential attribute of material substance, and all corporeal characteristics being modes of extension, the only changes a subject of corporeal characteristics is capable of undergoing are those that involve the replacement of one mode of extension with another; it is logically possible for something spherical to change into something cubical, but it would be logically impossible for something spherical to change into something having no shape whatever and no corporeal characteristics whatever. And it is not only Cartesians who find this principle compelling. Bernard Williams, a staunch anti-dualist, argues from the claim that "the understanding of what a given sort of thing is closely involves an understanding of under what determinables a thing of that sort *exemplifies determinates*" to the conclusion that "the possibility of disembodiment would show, not just that a person was a sort of thing that *did not necessarily* exemplify physical determinables, but that it was a sort of thing that *necessarily did not* exemplify such determinables."⁵ In that case, he says, "even embodied persons would not have physical attributes, but

5. Bernard Williams, "Are Persons Bodies?" in *Problems of the Self* (Cambridge: Cambridge University Press, 1973), p. 71.

would be nonphysical things associated with a body, i.e., a Cartesian account would apply." It certainly follows from this that if the dualist can establish (3) he is home clear.

Doubts can be raised about the essentialist premise, and I shall return to these later on. But it is, *prima facie*, a very plausible principle, and for now I shall assume that it is true.

In order to integrate the essentialist premise into the Jaeger-Malcolm argument, I shall take the liberty of substituting slightly different propositions for (2) and (3). The argument now goes as follows:⁶

- (1) *I think I am breathing* entails *I exist*. (Premise)
- (2') *I think I am breathing* does not entail *I have corporeal characteristics*. (Premise)
- (3') *I exist* does not entail *I have corporeal characteristics*. (From (1) and (2'))
- (4) Whatever has corporeal characteristics nonderivatively is essentially something having corporeal characteristics. (Premise)
- (5) I do not have corporeal characteristics nonderivatively. (From (3') and (4))

(3') follows from (1) and (2') for the same reason that our original (3) followed from (1) and (2); the form of the argument up to that point is the same. (4) is our essentialist premise. That (4) and (3') together entail (5) can be seen as follows. According to (3'), it is logically possible that I should exist without having corporeal characteristics. This means that I am not essentially something having corporeal characteristics. But if I am not essentially something having corporeal characteristics, then according to (4) I am not something having corporeal characteristics nonderivatively. And that is our conclusion.

I have resolved to leave premise (4) unchallenged for the time being. Premise (1) is a special case of Descartes's "Cogito," and would be regarded as incontestable by most philosophers, antidualists as well as dualists. If we had to drag our feet at (1)

6. I use the term "corporeal" in formulating the argument to give it an appropriate seventeenth-century flavor. When it suits my purposes I will sometimes use "physical" or "material" instead; nothing will turn on these changes in terminology.

in order to resist dualism, then at the very least the dualist would have won a very considerable moral victory. This leaves (2') as the premise to attack.

Before we consider how (2') might be attacked, we should notice that it seems on the face of it to be a very weak, even innocuous, claim. It merely asserts that it is not a *logical* consequence of my having a certain thought that I have corporeal characteristics. Putting it another way, it merely asserts that it is *logically* possible that I should have a certain thought without having any physical characteristics. Offhand one would have thought it unlikely that we could find incontestable, or at least highly plausible, premises which in conjunction with this seemingly very weak claim would entail a metaphysical doctrine as extreme as mind-body dualism. This is one of the things that makes the argument interesting.

But the central role played by (2') in our argument is interesting for another reason. (2') can be seen as an expression of the doctrine that mental states are "only contingently" connected with physical states of affairs, in particular bodily behavior. Among some recent philosophers, especially those influenced by the later philosophy of Wittgenstein, this doctrine is closely associated with Cartesian dualism. Wittgenstein remarks that "an 'inner process' stands in need of outer criteria," and his remarks about the status of behavior as criteria for the existence of mental states have been seen as an attack on Cartesianism and—what some have thought inseparable from this—as asserting a closer than contingent connection between mental states and behavior. A remark of Wittgenstein, noted by Malcolm, which seems directly in conflict with (2') is the following: "Only of a living human being and what resembles (behaves like) a living human being can one say: it has sensations; it sees; is blind; hears; is deaf; is conscious or unconscious."⁷ On the other hand, there are other philosophers, equally opposed to dualism, who think it a mistake to link the rejection of dualism with the acceptance of what I shall call the "conceptual connection thesis," namely the view that it in some way belongs to the concepts of the vari-

7. Ludwig Wittgenstein, *Philosophical Investigations*, ed. by G. E. M. Anscombe and R. Rhees, trans. G. E. M. Anscombe (Oxford: Basil Blackwell, 1953), para. 281.

ous mental states that any subject of these states must have a body in which, under appropriate conditions, the states can have *certain appropriate behavioral manifestations*. These philosophers (e.g., those proponents of the psychophysical identity theory who hold that the psychophysical identities are contingent) think that the falsity of dualism is a contingent matter, something that has been (or is in the process of being) discovered empirically, e.g., by the scientific developments mentioned earlier. And they see the conceptual connection thesis as a thinly disguised version of behaviorism, a doctrine they assume to have been discredited.

Now it might seem that a consideration of the Jaeger-Malcolm argument, and in particular the realization that (2') is apparently the premise to reject if one is to reject the argument, supports the former of these views—that which links the rejection of dualism with the acceptance of the conceptual connection thesis. To reject (2') one must assert that there is an entailment from the self-ascription of thoughts to the claim that one has some corporeal characteristics. It can easily appear that the only plausible account of how there could be such an entailment would be along the lines of a behaviorist or criteriological account of mental concepts, one that entails the conceptual connection thesis. This would also seem to vindicate those who have seen the issue of whether dualism is true as an *a priori* issue. For whether (2') is true depends on whether one proposition *entails* another; and it is natural to assume—or at any rate traditional to assume—that whether one proposition entails another is something to be established *a priori*. The Jaeger-Malcolm argument appears to be an *a priori* argument in favor of dualism. And if we can refute that argument by showing the falsity of (2'), then, it seems, we will have an *a priori* argument *against* dualism. If this is so, then the recent developments in biological science are far less central to the case against dualism than most people, or at any rate most nonphilosophers, tend to assume.

But as we shall see, it is in fact not necessary to claim that dualism can be refuted on *a priori* or conceptual grounds in order to refute the Jaeger-Malcolm argument. There is a refutation of it which is perfectly compatible with its being an empirical matter whether dualism is true. That this refutation can

be given does not imply that it is not *also* possible to refute dualism on a priori grounds. But it is useful to see that it is possible to refute the argument without giving an a priori argument for the necessary falsity or conceptual incoherence of dualism—and in particular that it is possible to do so without invoking some version of the conceptual connection thesis. If we think that we can reject dualism only by rejecting (2'), and that we can reject (2') only by accepting the conceptual connection thesis, then we may feel that we have to accept the conceptual connection thesis as the price of rejecting dualism. And then we will be under pressure to believe something false. For while I think that there is an important truth lying behind the conceptual connection thesis, I think that this truth is not such as to imply the falsity of dualism; so anyone who accepts the conceptual connection thesis in order to avoid accepting dualism will not only be accepting it for the wrong reason but will be accepting a mistaken version of it. Once we have seen that our argument for dualism can be refuted without the use of the conceptual connection thesis, the latter doctrine can be considered on its own merits, without our being under pressure to accept too strong a version of it.

Now let us proceed to an examination of (2'). (2') says that the proposition "I think that I am breathing" (call this "A") does not entail the proposition "I have corporeal characteristics" (call this "B"). Propositions A and B are both subject-predicate propositions, and both have the same subject—the grammatical subject being in both cases the word "I." Now when the question arises whether one such proposition entails another, what is usually in question is whether there is between them what I will call a "predicate entailment." There is a predicate entailment between two such propositions just in case they have the same subject and it is a necessary truth that whatever satisfies the predicate of the one also satisfies the predicate of the other. In other words, there is such an entailment from the proposition "a is F" to the proposition "a is G" just in case the proposition "Whatever is F is G" is a necessary truth. Thus there is a predicate entailment from "Jones is a bachelor" to "Jones is unmarried," since it is a necessary truth that whoever is a bachelor is unmarried, and there is a predicate entailment from "This fig-

ure is trilateral" to "This figure is triangular," since it is a necessary truth that whatever is trilateral is triangular.

If the universal necessary truth which backs up a particular entailment is analytically true, let us say that the predicate entailment is an analytic entailment. The entailment from "Jones is a bachelor" to "Jones is unmarried" is an analytic predicate entailment, if anything is. Many philosophers have thought that all predicate entailments are analytic. An even more common assumption has been that all such entailments are knowable *a priori*. Thus, consider the entailment from "This book is red" to "This book is not green," which holds in virtue of the necessary truth "Nothing is both red and green." Some philosophers have held that the latter proposition is synthetic rather than analytic; but even these philosophers have generally held that this proposition, or our knowledge of it, is *a priori*; it has in fact been offered as a non-Kantian example which supports the Kantian doctrine that there are synthetic *a priori* truths. If someone thinks that in order to refute the Jaeger-Malcolm argument one must be able to give an *a priori* disproof of dualism, this will be because he assumes that the entailment at issue in (2') would have to be a predicate entailment, and one that is analytic or at least *a priori*. And I think it is just this assumption that underlies the plausibility of (2'), i.e., the denial that there is an entailment from proposition A to proposition B. It is very plausible, indeed I think it is true, that there is no analytic predicate entailment from A to B; and the view that there is a predicate entailment which is *a priori* without being analytic does not seem more promising than the view that there is an analytic predicate entailment.

Recently the assumption that all predicate entailments are *a priori* has been called into question by Saul Kripke's contention that there are necessary truths about natural kinds that are *a posteriori*, i.e., empirical, rather than *a priori*. For example, according to Kripke the proposition "Gold is the element having the atomic number 79" is necessarily true, if true at all, even though it is not *a priori*; so assuming that it is true, the entailment from "My ring is made of gold" to "My ring is made of the element having atomic number 79" is a predicate entailment that is not *a priori*. So it would be possible for someone who

agrees that there is no a priori entailment from A to B to reject (2') on the grounds that there may be a predicate entailment from A to B that is not a priori.

But while this is a possible way of challenging the Jaeger-Malcolm argument, it is not the one I shall pursue. What I shall argue is that there may be an entailment from A to B which is not only not an a priori entailment but is not a predicate entailment at all, and that to assume without argument that there is not such an entailment amounts to begging the question in favor of dualism. Or at any rate, this is so on the assumption that premise (4) of the argument is true. That there are entailments between propositions like A and B that are not predicate entailments is another of the consequences of Kripke's work. But as we shall see, this claim is in fact something Descartes himself is committed to. Indeed, it is a consequence of one of the premises of the very Cartesian argument we are considering; it is implied by premise (4), the "essentialist premise."

Let us suppose that Kripke is right in thinking that it is necessarily true, if true at all, that Margaret Truman was born of Harry and Bess Truman. Or, more exactly, it is necessarily true that if Margaret exists (or ever existed) she was born of those parents. This follows from the plausible claim that in any possible world in which Margaret exists she was born of those parents; that however much a possible history of Margaret differs from her actual history, it cannot differ from it in starting with a birth from different parents. If this is so, the proposition "Margaret Truman sings" entails the proposition "Margaret Truman was born of Harry and Bess Truman"—assuming, of course, that the name "Margaret Truman" is used to refer to the right Margaret Truman. This is plainly not a predicate entailment; it is not necessarily true, or true at all, that whatever sings was born of Harry and Bess Truman. Yet it seems to be an entailment. Let us call such entailments "subject entailments." Where P and Q are subject-predicate propositions having the same subject, there is a subject entailment from P to Q if P entails the existence of its subject and Q ascribes to that subject some essential property.⁸

8. It might be better to speak of subject entailments and predicate entailments as holding between *statements*, where the identity of a statement de-

It is not necessary to maintain that there actually are any essential properties, and that there are any subject entailments that are not predicate entailments, in order to refute the Jaeger-Malcolm argument. For we can argue *ad hominem* by pointing out that any proponent of that argument is committed to there being such entailments. According to premise (4), anything having corporeal characteristics necessarily has corporeal characteristics. So consider the propositions expressed by "This is ten years old" and "This has corporeal characteristics," where the word "this" refers to something, let it be my watch, which has some corporeal characteristics. If (4) is true, the first proposition will entail the second; for according to (4) the thing in question must have corporeal properties in order to exist, and since it must exist in order to be ten years old, it follows that it must have corporeal characteristics in every world in which it is ten years old. And one proposition entails another just in case it is logically impossible for the one to be true without the other being true; or in other words, just in case the second is true in every possible world in which the first is true.

Many philosophers have found Descartes's First and Second Meditations a convincing proof that there can be no entailment from propositions like A to proposition B; for if one can coherently doubt the existence of a material world, and so of one's own body, while being unable to doubt one's own existence as a thinking being, it is natural to conclude that there can be no entailment from propositions like A (self-ascriptions of thoughts) to proposition B. A related argument for (2') is from the supposed imaginability of oneself, or someone, lacking corporeal characteristics. I think that this comes to much the same thing as arguing to the logical possibility of "I think but have no corporeal characteristics (or will at some time have no corporeal

depends on the sentence used to express it as well as on its propositional content. I would then want to stipulate that statements count as "having the same subject" only if they both have the same grammatical subject (contain the same singular referring expression) and the grammatical subject has the same referent in both. The statements I would express by saying "Hesperus is a planet" and "Phosphorus has physical characteristics" would not have the same subject in this sense, and there would be neither a subject entailment nor a predicate entailment between them.

characteristics)" from its epistemic possibility. But the most such arguments show is that there is no a priori predicate entailment from A to B, and we have now seen that establishing this is far from enough to establish (2'). To have a case for (2') we must have a case for thinking, not merely that there is no a priori predicate entailment from A to B, but also that there is no a posteriori subject entailment between them. But how could it be argued that *there is no subject entailment here?* Given that one accepts premise (4) of the argument, one could do so only by arguing that one does not possess, nonderivatively, any corporeal characteristics. But that is precisely the conclusion of our argument. So it appears that in order to establish the crucial premise of the argument one would have to establish its conclusion. The argument therefore appears to be question-begging.

In order to reject the Jaeger-Malcolm argument one need not claim to know that premise (2') is false; one need only claim that no good reason has yet been given for thinking that it is true—and in particular, that the fact (if it is one) that there is no predicate entailment from A to B is *no reason for thinking* this. Also, if one does in fact claim that (2') is false of oneself, one is not thereby committed to denying that there are creatures *for which it is true*. Whether there is a subject entailment between propositions expressed by sentences A and B will depend entirely on the essential nature of the creature referred to by the word "I" in these sentences; and it is at least *prima facie* conceivable that there should be creatures for which the entailment holds and others for which it does not. Perhaps dualism is true of the Martians but not of us. Or vice versa. There is a tendency to suppose that if the entailment holds for anyone it holds for everyone, and that if it fails for anyone it fails for everyone. If one thinks this, and also thinks that it is logically possible that there should be a creature which has thoughts but no body (perhaps because one thinks one can imagine becoming such a creature, or can imagine discovering that someone else has), one will think that the entailment fails for everyone and will thereby be committed to holding that dualism is true of all creatures having minds. If one makes the same assumption, but is convinced that dualism is false, one will think that one has to deny the logical

possibility of there being creatures that have thoughts without having bodies. But what leads to this assumption is the uncritical acceptance of the assumption that all entailments are predicate entailments. Once it is seen that this assumption is unfounded, or at any rate not available to anyone who accepts premise (4), it can be seen that the entailment at issue in (2') could hold for some creatures and not for others. And then it can be seen that it is at least *prima facie* consistent to hold that one is oneself an essentially corporeal being whose existence depends on its having corporeal characteristics, while holding that there may be other thinking beings which are capable of disembodied existence. It goes with this that whether a particular person is, nonderivatively, a subject of corporeal characteristics is a matter for empirical investigation, and not something to be discovered by *a priori* philosophical reflection.

III

It is possible that one source of the view that it is entirely an *a priori* matter whether dualism is true is the idea, which I have just argued to be mistaken, that in order to avoid dualism one must hold the view that for every mental state there is a predicate entailment from the having of that state to the having of some corporeal characteristics, together with the view that all predicate entailments must be *a priori*. But I am sure that this is not the only source. For I believe that there is an important version of dualism of which this view is true—a version such that the issue of whether it is true should be regarded as an *a priori* one. What partly accounts for the ambiguous status of dualism mentioned at the beginning of this paper, the fact that it is sometimes treated as an *a priori* thesis and sometimes as an empirical one, is that philosophers have not clearly distinguished the *a priori* version from a version which is, at least *prima facie*, empirical. What I say about these two versions of dualism here will be in part a summary of what I have said in more detail elsewhere; my excuse for going over this again is in part its bearing on the general theme of this paper, in part the fact that I

must do so in order to fulfill my promise to reexamine premise (4) of my version of the Jaeger-Malcolm argument.⁹

Let me begin by characterizing a position I will call "Minimal Dualism." According to this, for each person having mental states there is an incorporeal substance such that (a) what mental states the person has depends on what states the incorporeal substance has, (b) all causal connections involving mental states between the person's sensory input and behavioral output are mediated by states of this immaterial substance, and (c) it is possible for the person to exist, as a subject of mental states, without having a body, as long as the incorporeal substance exists and has the appropriate states. Now on one version of this view, the version I will call "Cartesian Dualism," the subject of mental states (i.e., the person) just is the incorporeal substance, and the states of the incorporeal substance just are the person's mental states. But notice that this is not implied by Minimal Dualism. It is compatible with Minimal Dualism, as I have characterized it, that the relation of the incorporeal substance and its states to the person and his mental states should be analogous to the relationship which a materialist thinks there is between a person's brain and its physiological states, on the one hand, and the person and his mental states, on the other. Most materialists would not want to say that a person is his brain; and whether or not the materialist says that the person's mental states are states of his brain, he will not say that all of the states of the person's brain are mental states. What our materialist will say is that the person's mental states are in some sense constituted by, or realized in, states of his brain. And on the version of Minimal Dualism which I will call "Non-Cartesian Dualism," the mental states of a person are constituted by, or realized in, the states of an incorporeal substance which can be thought of as a kind of ghostly brain. The situation in which a person exists in disembodied form, with the incorporeal substance separated from his body, will be analogous to the situation, envis-

9. See my "Immortality and Dualism" and "Postscript" in *Reason and Religion*, ed. by Stuart C. Brown (Ithaca: Cornell University Press, 1977), pp. 259-81 and 307-11. See also my "Embodiment and Behavior" in *The Identities of Persons*, ed. by Amelie Rorty (Berkeley and Los Angeles: University of California Press, 1976), pp. 109-37.

aged as possible by some materialists, in which the person exists as a brain in a vat.¹⁰

It will perhaps come as no surprise that the version of Minimal Dualism which seems to me an a priori thesis, and which I believe to be a priori false, is Cartesian Dualism, while the version which seems to me to be prima facie an empirical thesis, and empirically false, is Non-Cartesian Dualism. In the scope of the present paper I can do no more than sketch my reasons for thinking this.

Let me begin by trying to imagine a set of empirical observations which is as favorable to dualism as possible, and asking what sort of dualism, if any, these observations would support. I suppose that the observed phenomena should include lots of "spiritualistic" phenomena, apparent communications with the dead and the like, for which no physical explanations could be found. They should also include observations that indicate that neither the brain nor any other part of the body has the degree of complexity and the sort of organization it would have to have if its states were the sole causal basis of the enormously complex behavioral repertoires a person has in virtue of having the various mental states. It would also help if the observations were such as to give comfort to opponents of evolutionary theory, and supported the view that the gulf between the intellectual capacities of man and those of other animals is every bit as large as Descartes believed it to be.

Now if such a set of observations established dualism, what they would directly establish is the truth of what I have called *Minimal Dualism*—the view that we must postulate something over and above material things and physical processes in order to explain the observed phenomena we attribute to the mental states of human beings. Assuming that Non-Cartesian Dualism

10. The possibility of the position I have called "Non-Cartesian Dualism" was recognized by Hilary Putnam in "The Nature of Mental States" and "The Mental Life of Some Machines," when he pointed out that "the functional state hypothesis is not incompatible with dualism," and that a soul, or a system consisting of a body and a soul, could perfectly well be a Turing Machine or Probabilistic Automaton of the sort the "functional state hypothesis" holds minds to be. See Putnam, *Mind, Language and Reality*, *Philosophical Papers*, vol. 2 (Cambridge: Cambridge University Press, 1975), pp. 412 and 436.

is a coherent version of dualism, then in establishing Minimal Dualism these observations would not establish Cartesian Dualism. And it seems plain that there are no additional empirical observations that would tip the balance in favor of Cartesian Dualism over Non-Cartesian Dualism. Of course, so far the situation is symmetrical; if Cartesian Dualism is a coherent form of dualism, then in establishing Minimal Dualism these observations would not establish Non-Cartesian Dualism. But on further examination the symmetry breaks down.

Let us ask how Non-Cartesian Dualism could fail to be coherent. It would of course be incoherent if the very notion of an incorporeal substance turned out to be incoherent, or if it turned out that it is incoherent to suppose that material and incorporeal substances could interact causally. But in that case Cartesian Dualism would be incoherent as well, since *all* versions of Minimal Dualism would be incoherent. Let us suppose that this is not the case. Assuming, then, that the notion of an incorporeal substance is in order, and that causal interaction between material and immaterial substances is not a logical impossibility, it would seem that Non-Cartesian Dualism could fail to be coherent only if it were incoherent to suppose that mental states are related to states of incorporeal substances in the way materialists think that *mental states are related to physical states* of the brain. But since the notion of an incorporeal substance is a negative notion, i.e., is the notion of a substance that is *not* physical or material, such an incoherence could not be due to some special feature of this notion which precludes the possibility of *mental states being realized in, or constituted by, states of incorporeal substances*. Assuming, as we are, that there are coherent forms of dualism, it cannot belong to the concept of mental states that the only sorts of states they can be realized in, or constituted by, are physical states of material substances. So if, despite this, Non-Cartesian Dualism is not coherent, this can only be because it is not coherent to suppose that there are any states whatever that realize or constitute mental states in the way that materialists believe physical states of the brain realize or constitute them. But this is to say that Non-Cartesian Dualism will fail to be coherent only if *noneliminative materialism* fails to be coherent—where by “noneliminative materialism” I

mean the version of materialism which holds that there are mental states but that these are "nothing over and above" physical states. (Henceforth I shall take "noneliminative" as understood when I speak of materialism.) So anyone who thinks that it is an empirical issue whether materialism or dualism is true must think that Non-Cartesian Dualism is a coherent doctrine. There is no such argument to show that such a person must think that Cartesian Dualism is a coherent doctrine; and as I shall now indicate, I think there are reasons to hold that such a person ought to think that Cartesian Dualism is *not* a coherent doctrine.

A proponent of Non-Cartesian Dualism could hold that Non-Cartesian Dualism is true in the actual world but that there are other possible worlds in which materialism is true. For while he holds that *in fact mental states are realized in, or constituted by,* certain nonphysical states of incorporeal substances, it is compatible with this that these states could be realized physically—just as it is compatible with the view that certain mental states are in fact realized in certain physical states that they could be realized in different ones. Indeed, just as a materialist can hold that the same mental states can have one sort of physical realization among human Earthlings and another sort of physical realization among the denizens of some remote planet, so someone who holds that Non-Cartesian Dualism is true of himself and the rest of his kind might nevertheless allow that elsewhere in the universe there may be creatures psychologically like him whose mental states are realized physically, and are purely physical beings. A Cartesian Dualist, however, must hold that there is no possible world in which materialism is true, or in which there are purely physical beings which have mental states. According to him, *each mental property just is a certain nonphysical property*, and it is logically impossible that any such property should belong to a purely physical thing. If we could assume the traditional view that questions of logical possibility are always a priori, this would be enough to show that a proponent of Cartesian Dualism cannot regard the question of whether dualism or materialism is true as an empirical one. For such a dualist would have to think that the truth of materialism is logically impossible, and therefore not something it makes sense to investigate empirically. However, Kripke has complicated our

lives by showing that propositions whose truth or falsity is logically necessary can have the epistemological status of being a posteriori. So we need to carry the argument a bit further. In the present paper I can only indicate the direction in which the continuation of the argument would have to go.

Let me begin by formulating a doctrine, which I shall call "conceptual functionalism," which has recently been advanced by a number of philosophers.¹¹ This says, roughly, that the concept of a mind is the concept of a system of states that stand in certain relations, in particular causal relations, to one another and to behavior. What constitutes a particular state as being a particular mental state is its playing the role in such a system which is definitive of that mental state. For example, what constitutes a state as being a certain belief is the way in which it combines with certain desires, and with other beliefs, in the production and control of behavior directed at the satisfaction of those desires; perhaps, as a first approximation, it is the belief in a certain proposition because its contribution to behavior is such that it tends to maximize the satisfaction of the person's desires (whatever they might be) in circumstances in which that

11. Versions of this view can be found in David Armstrong, *A Materialist Theory of Mind* (London: Routledge & Kegan Paul, 1968); in David Lewis, "An Argument for the Identity Theory" in D. M. Rosenthal, ed., *Materialism and the Mind-Body Problem* (Englewood Cliffs, N.J.: Prentice-Hall, 1971), and "Psychophysical and Theoretical Identifications," *Australasian Journal of Philosophy*, 50 (1972); and in my own "Functionalism and Qualia," *Philosophical Studies*, 27 (1975):291-315, and "Some Varieties of Functionalism," *Philosophical Topics*, 12 (1981):93-119. A similar view can be found in Paul Grice, "Method in Philosophical Psychology (From the Banal to the Bizarre)," *Proceedings and Addresses of the American Philosophical Association*, 48 (1974-75):23-53. Other advocates of a functionalist approach to the philosophy of mind would deny the status of "conceptual truth" to both the functionalist position itself (the view that mental states are functional states) and to the functional definitions of particular states, and would instead assign to these the status of empirical hypotheses. See, for example, Hilary Putnam, "The Nature of Mental States." For an account of the different sorts of functionalism, and a criticism of functionalist theories, see Ned Block, "Troubles With Functionalism" in C. Wade Savage, ed., *Perception and Cognition: Issues in the Foundations of Psychology*, *Minnesota Studies in the Philosophy of Science*, vol. 9 (Minneapolis: University of Minnesota Press, 1978).

proposition is true.¹² As long as a state plays the appropriate role, it doesn't matter what it is like otherwise; thus it is that the same mental state can be "realized" in a variety of different ways. It is essential to this position, as I am characterizing it, that it is a conceptual truth that mental states are functional states, and that it is at least to some extent a conceptual matter what the functional definitions of particular sorts of mental states (*beliefs, desires, pains, etc.*) are. I also take it that conceptual functionalism and Cartesian Dualism are logically (and conceptually) incompatible. According to Cartesian Dualism, mental terms rigidly designate certain nonphysical properties and states, and all that belongs to their sense is that the properties and states thus designated are nonphysical. So if conceptual functionalism can be established by conceptual analysis, then Cartesian Dualism is not only false but conceptually false. This is what I believe to be the case. But in the present paper I have given no argument whatever in favor of conceptual functionalism. I have merely offered it as something which it is plausible to think can be established by conceptual analysis, and so a priori, and which rules out an interesting form of dualism.

It is worth observing that if conceptual functionalism is true, it can fairly claim to capture the germ of truth in the idea, mentioned earlier in connection with the Jaeger-Malcolm argument, that dualism is false, and conceptually false, *because* it denies the conceptual connections between mind and body. Conceptual functionalism is a cousin of what I earlier called the conceptual connection thesis; like the latter it makes the relations of mental states to bodily behavior partly definitive of them. It is not strong enough to warrant the claim that there is an analytic entailment between propositions A and B in premise (2') of the Jaeger-Malcolm argument, and so it is not strong enough to rule out all forms of dualism. In particular it does not rule out Non-Cartesian Dualism—for the latter can be seen as simply the view that the way in which mental states are in fact realized is

12. See Paul Grice, "Method in Philosophical Psychology (From the Banal to the Bizarre)."

in the incorporeal states of incorporeal substances. It is, however, strong enough to rule out Cartesian Dualism.

But now let us return to Non-Cartesian Dualism. If this is a coherent form of dualism, and perhaps (as I think) the only coherent form of dualism, it becomes necessary to reconsider premise (4) of the Jaeger-Malcolm argument (or, rather, of my reconstruction of that argument). According to Non-Cartesian Dualism, an embodied person is a system which is partly material and partly immaterial, the two parts interacting causally. The immaterial part is supposed to play the causal and functional role which materialists think is played by the brain. But if there could be such systems, why couldn't there also be systems in which what plays this causal and functional role is itself a system which is partly material and partly immaterial? In other words, if there can be persons animated by ghostly brains, why can't there be systems animated by partly ghostly brains? And if there could be such systems, why couldn't they undergo changes consisting in the replacement of material components with functionally equivalent immaterial ones, or vice versa?¹³ But if this is possible, then it would seem as if it should be possible for someone to start off as a completely physical being and to end up, as the result of a series of such replacements, having a completely immaterial substance playing the role of the brain. And if the continued existence and functioning of this ghostly brain could be sufficient for the existence of the person even if it were separated from his body (again, this may be compared with the case in which a brain is kept alive *in vitro*), then it would be true of something which at one time was a purely physical being, and had physical characteristics in a nonderivative way, that it could exist without having any physical properties at all. And if this is true then premise (4) of my reconstruction of the Jaeger-Malcolm argument is false. If so, then one cannot say, as I did earlier, that premise (2') of the argument begs the question. For if (4) is false, the assertion of (2') is not after all tantamount to the denial that one has corporeal characteristics in a nonderivative way. But this will be cold com-

13. Richard Boyd envisages a similar possibility in "What Physicalism Does Not Entail" in Ned Block, ed., *Readings in Philosophy of Psychology* (Cambridge, Mass.: Harvard University Press, 1980).

fort for the proponent of the Jaeger-Malcolm argument, for he can now be impaled on the horns of a dilemma. We can point out that on the assumption that (4) is true premise (2') begs the question, while on the assumption that (4) is false the argument is simply unsound.

It may be objected that if so-called immaterial substances could be functionally equivalent to material substances in the way just imagined, and if they could interact causally with them in ways analogous to those in which parts of the brain interact with one another and the rest of the body, then there could be no good reason for calling them "immaterial" or "nonphysical." I suspect that if this objection is pressed consistently it will deny the possibility of any sort of causal interaction between the material and the immaterial, or between the physical and the non-physical. The idea will be that whatever can interact causally with physical systems must itself be physical. In the end, I think, this will amount to a denial that the notion of an immaterial substance, a substance whose properties are nonphysical, is a coherent one. To argue this is to mount an a priori attack on dualism which is very different from the sorts of attack I have considered so far. This objection raises the difficult question of how, precisely, the terms "physical" and "material" are to be understood in formulations of the mind-body problem, and whether there is any acceptable understanding of them which makes dualism a coherent answer to the problem. I cannot pursue this question in any detail here, and will conclude with just a couple of brief remarks on it. First of all, the view that it follows from the correct analysis of the meaning of the term "physical" that whatever can interact causally with a physical system must itself be physical, and that any substance that interacts causally with human bodies is *ipso facto* a material substance having physical properties, is difficult to reconcile with the fact, mentioned earlier, that we can imagine a history of empirical observations which would seem to establish, or at least support, some version of dualism. If such observations actually occurred, a materialist who tried to use such an analysis to convince dualists that their views had not been vindicated would be accused, with some justice, of semantic hanky panky, to say nothing of being a sore loser. On the other hand, we certainly would not

want to limit the application of the term "physical" to phenomena that are reducible to phenomena *currently* recognized by physics, and I think that we have some tendency to so use that term that any entity or phenomenon which we have to posit in order to explain physical phenomena will itself count as physical. Perhaps there is some indeterminacy in our notion of the physical, and there is simply no fact of the matter as to whether certain imaginable happenings should count as the discovery of nonphysical, or immaterial, substances, or whether, on the other hand, they should count as the discovery of physical entities of a novel kind. If so, we have a further explanation, beyond those already given in this paper, of why the issue of dualism seems to be constantly shifting its status, presenting itself sometimes as an a priori issue and sometimes as an empirical one.

Books and Articles

by Norman Malcolm

Books

- Ludwig Wittgenstein: A Memoir*, with a biographical sketch by
G. H. von Wright (Oxford University Press: London, 1958).
Dreaming (Routledge & Kegan Paul: London, 1959).
Knowledge and Certainty (Prentice-Hall: Englewood Cliffs, N.J.,
1963).
Problems of Mind: Descartes to Wittgenstein (Harper and Row:
New York, 1971).
Memory and Mind (Cornell University Press: Ithaca, N.Y., 1977).
Thought and Knowledge (Cornell University Press: Ithaca, N.Y.,
1977).

Articles

- 1940: "Are Necessary Propositions Really Verbal?" *Mind*, 49,
no. 194, pp. 189-203.
"The Nature of Entailment," *Mind*, 49, no. 195, pp.
333-47.
1942: "Certainty and Empirical Statements," *Mind*, 51, no. 201,
pp. 18-46.
"Moore and Ordinary Language," in Paul Arthur Schilpp
(ed.), *The Philosophy of G. E. Moore*, *The Library of*
Living Philosophers (Northwestern University: Evans-
ton and Chicago).
1949: "Defending Common Sense," *The Philosophical Review*,
63, no. 2, pp. 201-21.

- 1950: "Russell's Human Knowledge," *The Philosophical Review*, 59, no. 1, pp. 94-106.
"The Verification Argument," in Max Black (ed.), *Philosophical Analysis* (Cornell University Press: Ithaca, N.Y.). Reprinted with revisions and additional footnotes in *Knowledge and Certainty*.
- 1951: "Philosophy for Philosophers" (correct title: "Philosophy and Ordinary Language"), *The Philosophical Review*, 60, no. 3, pp. 329-40.
- 1952: "Knowledge and Belief," *Mind*, 61, no. 242, pp. 178-89. Reprinted with revisions and additional footnotes in *Knowledge and Certainty*.
- 1953: "Direct Perception," *Philosophical Quarterly*, 3, no. 13, pp. 301-16. Reprinted with revisions and additional footnotes in *Knowledge and Certainty*.
"Moore's Use of 'Know,'" *Mind*, 62, no. 246, pp. 241-47.
- 1954: "On Knowledge and Belief," *Analysis*, 14, pp. 94-97.
"Wittgenstein's Philosophical Investigations," *The Philosophical Review*, 63, no. 4, pp. 530-59. Reprinted with corrections and additional footnotes in *Knowledge and Certainty*.
- 1956: "Dreaming and Scepticism," *The Philosophical Review*, 65, no. 1, pp. 14-37.
- 1957: "Dreaming and Scepticism: A Rejoinder," *Australasian Journal of Philosophy*, 35, pp. 201-11.
- 1958: "Knowledge of Other Minds," *The Journal of Philosophy*, 55, no. 23, pp. 969-78. Reprinted in *Knowledge and Certainty*.
- 1959: "Stern's Dreaming," *Analysis*, 20, no. 74, p. 47.
- 1960: "Anselm's Ontological Arguments," *The Philosophical Review*, 69, no. 1, pp. 41-60. Reprinted with new footnotes in *Knowledge and Certainty*.
- 1961: "Professor Ayer on Dreaming," *The Journal of Philosophy*, 58, no. 11, pp. 294-97.
- 1962: "George Edward Moore," *Ajatus*, 24. Finnish translation of a paper first published in English in *Knowledge and Certainty*.
"Memory and the Past," *The Monist*, 45, no. 2, pp.

- 247-66. Reprinted as one of "Three Lectures on Memory" in *Knowledge and Certainty*.
- 1963: "George Edward Moore," in *Knowledge and Certainty*.
"Three Lectures on Memory" ("Memory and the Past," "Three Forms of Memory," and "A Definition of Factual Memory"), in *Knowledge and Certainty*.
- 1964: "Is It a Religious Belief that 'God Exists?'" in John Hick, ed., *Faith and the Philosophers* (St. Martin's Press: New York).
"Scientific Materialism and the Identity Theory," *Dialogue*, 3, no. 2, pp. 115-25.
- 1965: "Descartes' Proof that His Essence Is Thinking," *The Philosophical Review*, 74, no. 3, pp. 315-38. Reprinted in *Thought and Knowledge*.
"Rejoinder to Mr. Sosa's 'Professor Malcolm on "Scientific Materialism and the Identity Theory,"'" *Dialogue*, 3, pp. 424-25.
- 1967: "Explaining Behavior," *The Philosophical Review*, 76, no. 1, pp. 97-104.
"The Privacy of Experience," in Avrum Stroll, ed., *Epistemology: New Essays in the Theory of Knowledge* (Harper and Row: New York). Reprinted in *Thought and Knowledge*.
"Wittgenstein, Ludwig Joseph Johann," in Paul Edwards, ed., *The Encyclopedia of Philosophy* (The Macmillan Company & The Free Press: New York), 5, pp. 327-40.
"Wittgenstein's *Philosophische Bemerkungen*," *The Philosophical Review*, 76, no. 2, pp. 220-29.
- 1968: "The Conceivability of Mechanism," *The Philosophical Review*, 77, no. 1, pp. 45-72.
- 1970: "Memory and Representation," *Noûs*, 4, no. 1, pp. 59-71.
"Wittgenstein and the Nature of Mind," *American Philosophical Quarterly Monograph*, no. 4 (Oxford). Reprinted in *Thought and Knowledge*.
- 1971: "The Myth of Cognitive Processes and Structures" in T. Mischel, ed., *Cognitive Development and Epistemology* (Academic Press: New York). Reprinted in *Thought and Knowledge*.

- 1972: "Ludwig Wittgenstein: Purity and Passion" in B. Mazlish, ed., *The Horizon Book of Makers of Modern Thought* (American Heritage: New York).
- 1973: "Thoughtless Brutes," Presidential Address, *Proceedings of the American Philosophical Association*, 46 (1972-73), pp. 5-20. Reprinted with revisions in *Thought and Knowledge*.
- 1974: "Behaviorism as a Philosophy of Psychology" in T. W. Wann, ed., *Behaviorism and Phenomenology: Contrasting Bases for Modern Psychology* (University of Chicago Press: Chicago). Reprinted in *Thought and Knowledge*.
- 1975: "Author's Response," part of an Author-Reviewer Symposium on *Problems of Mind: Descartes to Wittgenstein*, in *Philosophy Forum*, 14, pp. 289-306.
- "The Groundlessness of Belief" in Stuart Brown, ed., *Reason and Religion* (Cornell University Press: Ithaca, N.Y.). Reprinted in *Thought and Knowledge*.
- 1976: "Memory as Direct Awareness of the Past" in Godfrey Vesey, ed., *Impressions of Empiricism, Royal Institute of Philosophy Lecture, 1974-75* (St. Martin's Press: London).
- "Wittgenstein and Moore on the Sense of 'I know'" in Jaakko Hintikka, ed., *Essays on Wittgenstein in Honour of G. H. von Wright, Acta Philosophica Fennica*, 28, nos. 1-3, pp. 216-40. Reprinted with slight revisions in *Thought and Knowledge*.
- 1977: "Descartes' Proof that He Is Essentially a Non-Material Thing" in *Thought and Knowledge*.
- 1978: "Wittgenstein's Conception of First Person Psychological Sentences as 'Expressions,'" *Philosophical Exchange*, 2, pp. 59-72.
- 1980: "'Functionalism' in Philosophy of Psychology," *Proceedings of the Aristotelian Society*, n.s., 80 (1979-80), pp. 211-29.
- "Kripke on Heat and Sensation of Heat," *Philosophical Investigations*, 3, no. 1, pp. 12-20.
- 1981: "Kripke and the Standard Meter," *Philosophical Investigations*, 4, no. 1, pp. 19-24.

"Misunderstanding Wittgenstein," *Philosophical Investigations*, 4, no. 2, pp. 67-71.

"The Relation of Language to Instinctive Behavior," J. R. Jones Memorial Lecture, University College of Swansea.

1982: "Wittgenstein and Idealism," in Godfrey Vesey, ed., *Idealism Past and Present*, Royal Institute of Philosophy Lecture Series: 13, Supplement to *Philosophy* 1982 (Cambridge University Press: Cambridge, England).

INDEX OF NAMES

Adams, Robert M., 221n
 Ambrose, Alice, 12n
 Ancombe, G.E.M., 14n, 140, 211n,
 242n
 Anselm, 260
 Armstrong, David, 254n
 Atlas, Jay, 84n
 Ayer, A.J., 260

Bach-y-Rita, Paul, 144, 145n
 Baier, Annette, 63n
 Baker, Doris, 63n
 Baker, Gordon, 106n
 Baldwin, James, 22n
 Barker, Stephen F., 136n
 Beauchamp, Tom L., 136n
 Bennett, Jonathan, 69
 Berkeley, 130-31, 132n, 133-34, 158
 Berry, A.J., 163n
 Black, Max, 260
 Blake, William, 115
 Bleuler, E., 81n
 Block, Ned, 254n, 256n
 Bohm, David, 187
 Boleyn, Anne, 184
 Bonjour, Lawrence, 41n
 Boyd, Richard, 256n
 Braude, Stephen, 210n
 Brown, Stuart C., 250n, 262
 Bruner, Jerome, 197, 200, 210
 Burge, Tyler, 84n
 Burnyeat, Myles, 63n
 Butler, R.J., 127n, 149n

Castañeda, Hector-Neri, 215n
 Catherine of Aragon, 184
 Cavendish, Henry, 163
 Cheseldon, Mr., 134
 Chomsky, Noam, 200-202, 205, 210
 Cicero, 88n, 89
 Clement, W.C., 137n
 Collins, Carter C., 144
 Columbus, 107
 Cook, Monte, 91n

Daniels, Norman, 135n, 136n
 Davidson, Donald, 66, 74, 84n, 105n
 Davis, John W., 132n
 Deleuze, Gilles, 63
 Dennett, Daniel C., 227n
 Derrida, Jacques, 66n
 Descartes, 7, 9, 13-15, 17-19, 21, 24,
 28, 65-68, 70-71, 80, 82, 233, 235,
 238-41, 246-47, 250-53, 255-56,
 259, 261-62
 Diamond, Cora, 211n
 Disney, Walt, 186
 Donnellan, Keith, 221
 Duggan, Timothy J., 135n, 139n, 140
 Dummett, Michael, 106n

Eccles, John, 236n
 Edwards, Paul, 261

Firth, Roderick, 43n
 Fodor, Jerry, 191-204, 206-11

- Foucault, Michel, 66n, 80
 Frankfurt, Harry, 66n, 68, 75
 Frege, Gottlob, 160, 169-70
- Geach, Peter, 14n
 Goldberg, Bruce, 206n, 210
 Gombay, André, 66n
 Grandy, Richard, 72
 Grice, H.P., 30, 130, 149-50, 155, 254n, 255n
 Guarniero, Gerard, 147n
 Gunderson, Keith, 195, 210
- Hacker, P.M.S., 106n
 Hamilton, William, 135n, 136n, 144
 Harman, Gilbert, 84n, 105n
 Henry VIII, 184
 Hercules, 153-54
 Hintikka, Jaakko, 29n, 91n, 262
 Hughes, Sally Smith, 119, 120n, 121
- Immerwahr, John, 136n
- Jaeger, Robert, 234-38, 241, 243, 247-48, 250, 257
 James, William, 30
 Jaspers, Karl, 81
 Jeans, James H., 152n
 Jones, J.R., 263
- Kant, 245
 Kaplan, David, 84n, 91, 215n
 Kenny, Anthony, 66n, 67
 Kripke, Saul, 84-85, 87-96, 98, 105-14, 116-21, 123, 129, 229n, 245-46, 253, 262
- Laing, R.D., 63
 Lazerowitz, Morris, 12n
 Lehrer, Keith, 44n
 Leibniz, 42, 152n
 Lewis, David, 215n, 254n
 Locke, John, 87, 93-95, 98, 102-3, 131-32, 147, 152n
- Mackie, J. L., 132n
 Malcolm, Norman, 29n, 64-65, 77-78, 84n, 106, 129, 140n, 200n, 206n, 210-11, 234-38, 241-43, 247-48, 250, 257, 259, 261
 Maxwell, James Clerk, 152n
 Mill, J.S., 90
 Mischel, Theodore, 261
 Molyneux, William, 130-33, 147-48, 157
 Monck, W.H.S., 132n
 Moore, G.E., 3-13, 15, 19-25, 29n, 260-62
 Moravcsik, J.M.E., 91n
 Morgan, Michael J., 145n, 146, 147n
 Muirhead, J.H., 3n
 Munitz, M.K., 84n
 Murphy, Arthur, 7
- Noonan, Harold W., 152-54
- Park, Désirée, 132n
 Penfield, Wilder, 236n
 Perry, John, 215n
 Pinfold, Gilbert, 65
 Pitcher, George, 134n
 Popper, Karl R., 236n
 Price, H.H., 165-66
 Putnam, Hilary, 84-85, 87, 89-98, 102-4, 127-28, 209n, 210, 251n, 254n
- Reid, Thomas, 130, 133-41, 143-44
 Rhees, Rush, 242n
 Richardson, John T.E., 106n
 Rorty, Amelie, 250n
 Rorty, Richard, 233, 234n
 Rosenthal, D.M., 254n
 Roxbee Cox, J.W., 149n
 Russell, Bertrand, 11-13, 17, 260
- Savage, C. Wade, 254n
 Scheffler, Samuel, 232n
 Schilpp, Paul, 7n
 Schneider, K., 81n
 Searle, John, 111, 232n
 Shoemaker, Sydney, 49n, 106n
 Sluga, Hans, 230n

- Sorel, Thomas, 232n
 Sosa, Ernest, 261
 Starr, Ringo, 217
 Stern, Kenneth, 260
 Stewart, Dugald, 143-44
 Stroll, Avrum, 261
 Suppes, Patrick, 91n
 Szasz, Thomas, 63

 Taylor, J., 146n
 Tebaldi, David A., 136n
 Teichman, Jenny, 211n
 Thomson, Judith Jarvis, 130, 147-48,
 151n, 155
 Truman, Bess, 246
 Truman, Harry, 246
 Truman, Margaret, 246
 Tully, *see* Cicero
 Turing, A., 192-93, 210

 Velleman, David, 232n
 Vesey, Godfrey, 262-63
 von Mises, Richard, 188
 von Wright, G.H., 29n, 259, 262

 Wann, T.W., 262
 Wefald, Eric, 232n
 Wiggins, David, 152n
 Williams, Bernard, 78, 240
 Williams, Michael, 30n
 Wilwerding, Jon, 84n
 Winkler, W., 81n
 Wittgenstein, Ludwig, 8, 29n, 69-70,
 78, 106, 113-14, 123-26, 128-29,
 140n, 159-60, 170, 197, 200n, 210,
 227n, 242, 259-63

 Zemach, Eddy, 108n